Extending the Knowledge Base of Foresight: The Contribution of Text Mining

vorgelegt von
Victoria Kayser, M. Sc.
geb. in Böblingen

von der Fakultät VII – Wirtschaft und Management
der Technischen Universität Berlin
zur Erlangung des akademischen Grades

Doktor der Wirtschaftswissenschaften
- Dr. rer. oec. -

genehmigte Dissertation

Promotionsausschuss:

Vorsitzender: Prof. Dr. Jan Kratzer
Gutachter: Prof. Dr. Knut Blind
Gutachter: Prof. Dr. Carsten Dreher

Tag der wissenschaftlichen Aussprache: 24. Februar 2016

Berlin 2016

Abstract

The future is shaped and influenced by decisions made today. These decisions need to be made on a solid ground and diverse information sources should be considered in the decision process. For exploring different futures, foresight offers a wide range of methods for gaining insights. The starting point of this thesis is the observation that recent foresight methods particularly use patent and publication data or rely on expert opinion, but few other data sources are used. In times of big data, many other options exist and, for example, social media or websites are currently not a major part of these deliberations. While the volume of data from heterogeneous sources grows considerably, foresight and its methods rarely benefit from such available data. One attempt to access and systematically examine this data is text mining that processes textual data in a largely automated manner. Therefore, this thesis addresses the contribution of text mining and further textual data sources for foresight and its methods. After clarifying the potential of combining text mining and foresight, four concrete examples are outlined. As the results show, the existing foresight methods are improved as exemplified by roadmapping and scenario development. By exploiting new data sources (e.g., Twitter and web mining), new options evolve for analyzing data. Thus, more actors and views are integrated, and more emphasis is laid on analyzing social changes. Summarized, using text mining enhances the detection and examination of emerging topics and technologies by extending the knowledge base of foresight. Hence, new foresight applications can be designed. And, in particular, text mining is promising for explorative approaches that require a solid base for reflecting on possible futures.

Zusammenfassung

Die Zukunft wird von heutigen Entscheidungen geformt und beeinflusst. Diese Entscheidungen sollten auf einer soliden Basis getroffen werden sowie diverse Informationsquellen im Entscheidungsprozess in Betracht gezogen werden. Um verschiedene Zukünfte zu erkunden, bietet Foresight eine große Spannbreite an Methoden um neue Erkenntnisse zu gewinnen. Der Ausgangspunkt für diese Dissertation ist die Beobachtung, dass derzeitige Foresight-Methoden vor allem Patent- und Publikationsdaten nutzen oder sich auf Experteneinschätzungen stützen, aber wenig andere Datenquellen verwendet werden. Im Zeitalter von Big Data existieren viele andere Optionen und viele Textquellen, wie zum Beispiel soziale Medien oder Webseiten, sind derzeit kein Kernbestandteil dieser Überlegungen. Während das Datenvolumen aus heterogenen Quellen erheblich steigt, machen sich Foresight und seine Methoden das nicht zu nutzen. Ein Ansatz diese Daten systematisch zu erschließen und zu erforschen ist Text Mining, womit Textdaten weitestgehend automatisch verarbeitet werden. Deshalb adressiert diese Dissertation den Beitrag von Text Mining und weiterer Datenquellen zu Foresight und seinen Methoden. Nach einer grundsätzlichen Klärung des Potentials einer Kombination von Text Mining und Foresight, werden vier konkrete Beispiele vorgestellt. Wie die Ergebnisse zeigen, werden die bestehenden Foresight-Methoden verbessert wie für Roadmapping und Szenarioentwicklung veranschaulicht wird. Durch die Nutzung neuer Datenquellen (z. B.: Twitter und Web Mining) entstehen neue Möglichkeiten in der Datenanalyse. Dadurch können mehr Akteure und Sichtweisen integriert und die Analyse gesellschaftlicher Veränderungen stärker betont werden. Zusammengefasst verbessert Text Mining die Erkennung und Untersuchung von aufkommenden Themen und Technologien, indem die Wissensbasis von Foresight erweitert wird. Neue Foresight-Anwendungen können daraus entwickelt werden. Und besonders vielversprechend ist Text Mining für explorative Ansätze, die eine solide Basis erfordern, um Überlegungen über mögliche Zukünfte anzustellen.

Content

Figures

Tables

I. Introduction

The future is unknown and unpredictable but foresight offers ways for its exploration and estimation (e.g., Martin, 1995; Slaughter, 1995). In this context, a wide range of methods have evolved how to systematically look into the future and gain insights on future developments (e.g., Cuhls, 2008; Popper and Butter, 2008). In parallel, the volume of data from heterogeneous sources, especially on the web, considerably grows (e.g., Ortner et al., 2014) and the scientific output is constantly increasing (see e.g., Bornmann and Mutz, 2014). Currently, foresight and its methodology rarely benefit from this available data and its contribution is not explored. Thereby, relevant information sources are left out whereas this data could be used to perceive ongoing changes and make more precise statements about possible future developments and emerging technologies.

One attempt to access and systematically examine this data is text mining (Berry, 2004; Feldman and Sanger, 2008). Text mining processes textual data such as reports, blog entries, or Twitter data. From this, terms are extracted and analyzed for patterns and dependencies (e.g., Manning et al., 2009).

Concerning foresight methods, applications using text mining exist for patent- and publications analysis and some for roadmapping. But for most other foresight methods no effort has been spent on using text mining so far. For the analysis of technical developments, patents and scientific publications are analyzed in foresight for long (e.g., Tseng et al., 2007; Delen and Crossland, 2008). In contrast, social media data is rarely analyzed (Glassey, 2012; Yoon, 2012) and web data is only considered for desk research. Generally, the user-generated content on the web may be interesting for foresight to examine social perspectives and the user's perception of current developments. In addition, applications that compare or match textual datasets are rare. However, many options exist such as automatic data gathering and aggregation.

Arising opportunities for foresight from text mining are, for example, the exploitation of data sources not used so far to improve foresight results and existing foresight methods. This relates to examining currently neglected data sources as Twitter, newspapers, or websites. On the other hand, new approaches for data analysis and retrieval can be applied such as web mining. So the questions addressed in this thesis are how and what foresight benefits from text mining. In particular, this thesis explores how to enrich explorative foresight approaches by extending the knowledge base of foresight by additional data and stakeholder views. Therefore, concrete realizations are implemented where text mining is built in foresight methods such as roadmapping (Möhrle et al., 2013) or scenario development (Reibnitz, 1991). Thereby, it is expected to enhance the detection and examination of emerging themes and technologies for a solid base for decision making.

The main part of this thesis consists of five articles which combine different foresight methods, textual data sources, text mining approaches and scopes of foresight. The first article lays the conceptual framework and introduces foresight and text mining (see Section 1). This article describes different data sources and summarizes the state-of-the-art of recent combinations of foresight and text mining. The principal relevancy and added value of text mining for foresight is elaborated along the process of text mining respectively foresight. This article concludes with acknowledging the potential of text mining for foresight. In the following work, concrete applications are outlined.

The second article is a first methodological draft of how to combine foresight and text mining (see Section 2). The contribution lays in the systematic integration of text mining in technology roadmapping illustrated on the example of *cloud computing*. This article builds on scientific publication data and implements a first text mining approach. Concerning the data analysis, the focus lays on processing abstracts as an important preparatory step for analyzing longer texts such as reports in the following work.

The third article is located in the innovation system framework and addresses the role of media therein (see Section 3). When analyzing the state-of-the-art and related studies, obviously not much work compares different textual datasets. Therefore, this article develops an approach to automatically compare science and media reporting based on scientific abstracts and news reporting. Furthermore, the difference between content analysis and text mining is addressed.

The fourth article addresses the use of Twitter data in foresight and shows how *new* information channels contribute to foresight (see Section 4). For the common data sources in foresight (e.g., patents, publications), strength and weaknesses are explored and defined (e.g., Bonino et al., 2010; Cunningham et al., 2006). These limitations are less clear for other data sources, but require some basic considerations before they are used in foresight. This article considers different applications and use cases to reveal how to use Twitter in foresight exercises for both, the monitoring of topics and technologies, but also the active user engagement.

Building on the results of the fourth article, the fifth article proposes a new scenario process that uses web mining to capture the state-of-the-art (see Section 5). Links are extracted from Twitter data for systematic data retrieval. Concerning social media and web mining, the automatic analysis of large text volumes enables new opportunities for foresight and its methods.

In the final part, the results are summarized. Here, text mining for foresight is assessed and the implications for foresight are described. Then, a conclusion is drawn that outlines directions for future research.

References

Berry, Michael W. *Survey of text mining: Clustering, classification, and retrieval*. New York: Springer, 2004.

Bonino, Dario; Ciaramella, Alberto; Corno, Fulvio. “Review of the state-of-the-art in patent information and forthcoming evolutions in intelligent patent informatics.” *World Patent Information* 32, no. 1 (2010): 30–38.

Bornmann, Lutz; Mutz, Rüdiger. “Growth rates of modern science: A bibliometric analysis.” *CoRR* abs/1402.4578 (2014).

Cuhls, Kerstin. *Methoden der Technikvorausschau - eine internationale Übersicht*. Stuttgart: IRB Verlag, 2008.

Cunningham, Scott W; Porter, Alan L; Newman, Nils C. “Special issue on tech mining: Tech Mining: Exploiting Science and Technology Information Resources.” *Technological Forecasting and Social Change* 73, no. 8 (2006): 915–922.

Delen, Dursun; Crossland, Martin D. “Seeding the survey and analysis of research literature with text mining.” *Expert Systems with Applications* 34, no. 3 (2008): 1707–1720.

Feldman, Ronen; Sanger, James. *The text mining handbook: Advanced approaches in analyzing unstructured data*. Cambridge, New York: Cambridge University Press, 2008.

Glassey, Olivier. “Folksonomies: Spontaneous crowd sourcing with online early detection potential?” *Futures* 44, no. 3 (2012): 257–264.

Manning, Christopher D; Raghavan, Prabhakar; Schütze, Hinrich. *An Introduction to Information Retrieval*. New York: Cambridge University Press, 2009.

Martin, Ben R. “Foresight in science and technology.” *Technology Analysis & Strategic Management* 7, no. 2 (1995): 139–168.

Möhrle, Martin G., Ralf Isenmann, and Robert Phaal, eds. *Technology roadmapping for strategy and innovation: Charting the route to success*. Berlin [et al.]: Springer, 2013.

Ortner, Heike; Pfurtscheller, Daniel; Rizzolli, Michaela; Wiesinger, Andreas. “Zur Einführung – Datenflut und Informationskanäle.” In *Datenflut und Informationskanäle*. 1st ed., edited by Heike Ortner, Daniel Pfurtscheller, Michaela Rizzolli and Andreas Wiesinger. Innsbruck: Innsbruck Univ. Press, 2014.

Popper, Rafael; Butter, Maurits. “How are foresight methods selected?” *foresight* 10, no. 6 (2008): 62–89.

Reibnitz, Ute. *Szenario-Technik: Instrumente für die unternehmerische und persönliche Erfolgsplanung*. Wiesbaden: Gabler, 1991.

Slaughter, Richard. *The foresight principle: Cultural recovery in the 21st century*. London, England: Adamantine Press; Adamantine, 1995.

Tseng, Yuen-Hsien; Lin, Chi-Jen; Lin, Yu-I. “Text mining techniques for patent analysis.” *Information Processing & Management* 43, no. 5 (2007): 1216–1247.

Yoon, Janghyeok. “Detecting weak signals for long-term business opportunities using text mining of web news.” *Expert Systems with Applications* 39, no. 16 (2012): 12543–12550.

Publication and Submission Record

This thesis consists of five articles. The co-authors have kindly given their consent that these articles are used in this book.

"*The Potential of Text Mining for Foresight",* together with parts of Section III (Conclusion), is submitted to *Technological Forecasting and Social Change.*

The second article is a revised version of "*Text Mining for Technology Roadmapping: The strategic Value of Information*", co-authored by Kerstin Goluchowicz and Antje Bierwisch. It was presented at the ISPIM 2013 in Melbourne and honored with the *Alex Gofman Best Student Paper Award*. It is published in the *International Journal of Innovation Management*[1].

"*The Role of Media in the Innovation System: Measuring the Knowledge Flow between Science and Public*" is currently under review in *Technological Forecasting and Social Change*. Earlier versions were presented at the *EuSPri-Conference 2015* in Helsinki and at the *EuSPri-Summer School 2015* in Utrecht.

"*Using Twitter for Foresight: An opportunity?*" is co-authored by Antje Bierwisch. This article was presented at the *ISPIM 2015* in Budapest and is currently under review in *Futures*.

"*Web-based Scenario Development: Process Improvements*" is co-authored by Erduana Shala. This article will be presented at the *Scenario 2015*-Conference in Warwick and is submitted to *Technological Forecasting and Social Change*.

[1] *Electronic version of an article published as International Journal of Innovation Management, Vol. 18, No 03, 2014,* 1440004; Doi: 10.1142/S1363919614400040 © *World Scientific Publishing Company http://www.worldscientific.com/worldscinet/ijim*

Acknowledgements

I would like to thank Knut Blind for his support and supervision and Carsten Dreher for reviewing my thesis.

My thesis profited a lot from the opportunity to present at conferences, the CiF Colloquium, and the Summer Schools I attended. Thanks for the feedback and comments I received.

This thesis was realized during my time at the Fraunhofer ISI in Karlsruhe. First, I would like to thank Antje Bierwisch for her comments on my work. She is one of those who decisively contributed to the completion of this thesis. Thanks are also due to Kerstin Cuhls, particularly for her advice on getting my work published. Then, I want to express my special thanks to my PhD-colleague Erduana Shala, for the constructive research dialogue we had over the last three years and the productive atmosphere in our office.

Furthermore, I would like to express my gratitude to Jonas Prior for his patience and the technical advice on the many programming challenges I had during my work on this thesis. And I would like to thank my family and friends.

Victoria Kayser

March 2016

II. Five Articles on Text Mining for Foresight

1 The Potential of Text Mining for Foresight

Abstract: At present time, we are faced with a growing volume of (textual) data. Currently, this increase in data is not met in foresight and its methodology. Principally, text mining offers ways to systematically access and analyze this data. In a largely automated manner, it may aggregate and structure thematic information and support foresight processes. At the intersection of foresight and text mining, currently not much work exists and opportunities are missed. Therefore, the objective of this article is to explore the potential of text mining for foresight considering different data sources, text mining approaches and foresight methods. Examples are outlined on roadmapping and scenario development. As the results show text mining is most promising to improve foresight methods and exploit further data sources.

Keywords: Foresight, Text Mining, Innovation, Social Media, Data Analysis, Roadmapping, Scenario Development

1.1 Introduction

Foresight offers ways for exploring and estimating the future. In this context, a wide range of methods have evolved for looking into the future and gain insights on possible future developments (e.g., Popper and Butter, 2008; Martin, 1995). Nowadays, the volume of data from heterogeneous sources considerably grows (Ortner et al., 2014) and the scientific output is constantly increasing (see, e.g., Bornmann and Mutz, 2014). In foresight, this causes a challenge in identifying the relevant data from the huge quantity of available information (see e.g., Montoyo et al., 2012). This increase in data is currently not met in foresight and its methods does not benefit from the available data. Nevertheless, this data may be used to perceive ongoing changes and make statements about future developments and emerging technologies.

Text mining is an approach to analyze textual data (Feldman, Sanger 2006; Manning et al. 2009). It extracts the most relevant terms from texts and analyzes them with predefined methods. Thereby, further data sources are accessible by integrating text mining in foresight, especially unstructured and large datasets, to be considered in a comprehensive way.

In an explorative case, foresight practices build upon the available data about the subject matter, e.g. literature, patents or bibliometric data. However, at the intersection of foresight and text mining, currently not much work exists, except for patent and publication analysis. For the analysis of technical developments, patents and scientific publications are analyzed in foresight for long (Kostoff, 2012; Abbas et al., 2014). However, the scope of foresight not only rests on identifying new technologies and current trends in different manner but comprises societal challenges (Salo and Cuhls, 2003). Generally, the user-generated content on the web may be interesting for foresight to examine social perspectives. However, the web data is, so far, rarely considered for a systematic examination (Yoon, 2012; Cachia et al., 2007; Glassey, 2012).

Apart from a fundamental consideration of the two building blocks text mining and foresight, the objective of this paper is to argue the potential of text mining for foresight and its methods. The angles of this article are the abstraction level of foresight (*micro* to *meso*), the diversity of foresight methods, different text mining approaches and the variety of textual data sources. Thereby, one aspect is to consider potential data sources that may be analyzed in the future to improve foresight results. Furthermore, it is examined to which extent foresight and its methods can be improved by the results of text mining. Therefore, this article addresses how and what foresight benefits from text mining.

This article begins with the fundamentals of foresight and the basic principles of text mining in Section 1.2. Then, Section 1.3 addresses the use of text mining for foresight. Different data sources are described, the state-of-the-art concerning existing implementations is summarized, and possible future applications are outlined. Finally, the results are discussed in the framework of foresight and a conclusion is drawn in Section 1.4.

1.2 The two Components: Foresight and Text Mining

The following section introduces the two main components of this thesis, foresight and text mining and gives an overview on the recent debate.

1.2.1 Foresight

In general, foresight is a systematic process to look 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 – so there is not the one future. Principally, the future cannot be predicted but is shaped by decisions and actions made today. Foresight serves for assessing the consequences and implications of present actions, early warning, and thinking about desirable futures and implications of possible future events. So, it is an action-oriented decision support by bringing together the relevant stakeholders for an open discourse about possible futures.

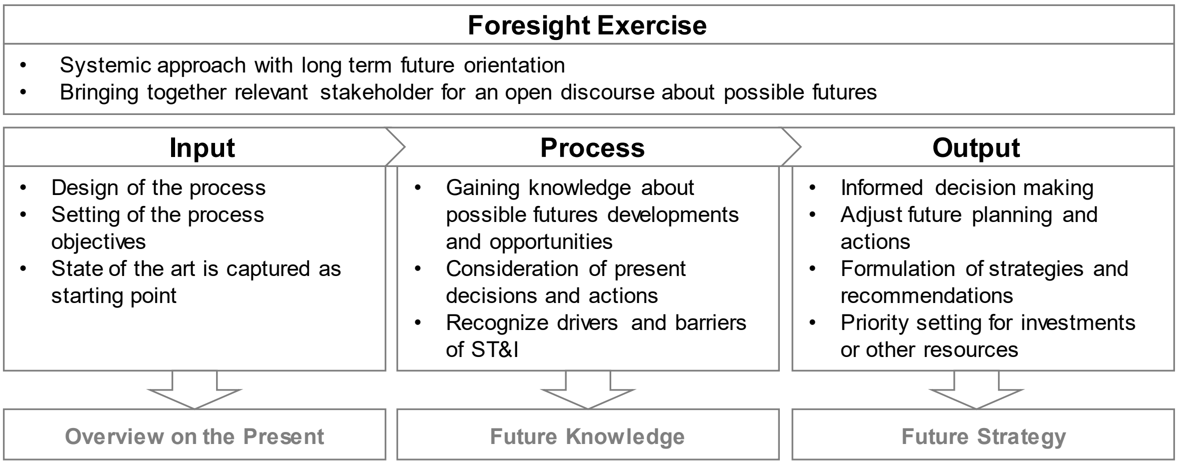
In general, foresight builds on of a set of different methods (e.g., Popper and Butter, 2008) such as roadmapping (Barker and Smith, 1995; Möhrle et al., 2013) or scenario development (Reibnitz, 1991; van der Heijden, 2005). Which set of methods to apply depends on the scope and focus of the foresight 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 thesis due to their commonalities.

Foresight is not fully structured (Bañuls and Salmeron, 2011) but often follows a certain order and is a sequence of steps (see, e.g., Martin, 1995; Horton, 1999; Voros, 2003; Da Costa et al., 2008; de Miranda Santo et al., 2006). This underlines the modular character of foresight: Depending on the objectives and application level, different methods and tasks are combined. Building on these previous studies, foresight exercises might be aggregated to three phases as illustrated in Figure 1-1: input, process and output.

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

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

Output: The results are assessed, priorities are set and strategies are formulated (de Miranda Santo et al., 2006). This phase is about taking actions (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, politics or governmental actors expect recommendations for planning or setting priorities for research programs (e.g., Salo and Cuhls, 2003; Havas et al., 2010).



Foresight Exercise

- Systemic approach with long term future orientation
- Bringing together relevant stakeholder for an open discourse about possible futures

Input

- Design of the process
- Setting of the process objectives
- State of the art is captured as starting point

Process

- Gaining knowledge about possible futures developments and opportunities
- Consideration of present decisions and actions
- Recognize drivers and barriers of ST&I

Output

- Informed decision making
- Adjust future planning and actions
- Formulation of strategies and recommendations
- Priority setting for investments or other resources

Overview on the Present

Future Knowledge

Future Strategy

Figure 1-1 The process of foresight

Considering the larger framework, foresight is conducted with different scopes, for different areas (international, national, or regional), and on different abstraction levels (micro to meso). In general, foresight is very interdisciplinary (see, e.g., Hines and Gold, 2013), misses a clear theoretical basis (Fuller and Loogma, 2009; Öner, 2010) and is more a well-established field of practice (e.g., Andersen and Andersen, 2014). However, the present understanding of foresight has some (theoretical) influences, particularly from technology forecasting (see, e.g., Martino, 1993) and futures studies (Bell, 2003).

In the literature, innovation is often related to foresight (e.g., Linstone, 2011; Watts and Porter, 1997), but innovation (studies) is rarely considered as a theoretical fundament for foresight (Andersen and Andersen, 2014). Principally, the range of innovation from *micro* to *meso* level fits very well. Like foresight, innovation is considered on different angles reaching from the process view of innovation management to the systemic view.

Foresight and innovation management have some commonalities, particularly in the early phase of innovation (Cuhls, 2011). While foresight delivers input for the strategic orientation and guidance, innovation management deals with the process from idea generation and invention to market entry and its aims are more concrete (see, e.g., Tidd and Bessant, 2009). Foresight can contribute to the innovation capacity of a firm by strategically exploring new business fields, initiating and contributing new ideas, and challenge the current processes (e.g., Rohrbeck and Gemünden, 2011). Principally, the unknown future is linked to business risks and a central element of a successful future strategy is the early recognition of trends and developments in the firm's environment to adapt to these recognized changes (e.g., Eisenhardt and Martin, 2000). Here, foresight supports monitoring and scanning the environment to capture the *big picture*. Like innovation, foresight has to take framework conditions into account such as policy, regulation, or human capital and skills. So future developments are considered in a systemic framework (Martin and Johnston, 1999; Andersen and Andersen, 2014). For mapping the systemic framework, anticipating current developments and integrating an objective external view, approaches as text mining might be used as will be introduced in the following.

1.2.2 Text Mining

Due to the increasing volume of data from heterogeneous sources, the effort increases to overlook thematic fields and developments and to read the amounts of published studies and literature (Ortner et al., 2014). Techniques are necessary to identify the relevant data from the huge quantity of available information and then process it into knowledge to be used in decision making (Montoyo et al., 2012). For this purpose, text mining offers methods to accesses and analyze these textual data sources (Weiss, 2010; Feldman and Sanger, 2008). Text mining processes unstructured textual data to a structured format for further analysis. Typical tasks in text mining are, for example, the identification of clusters, frequencies or associations. An overview on different text mining applications is, for example, given in Miner (2012). Principally, text mining can be summarized in three steps as indicated in Figure 1-2. First, a data source is selected. Then, this data is preprocessed (step 2) and analyzed (step 3). Finally, the results require interpretation.

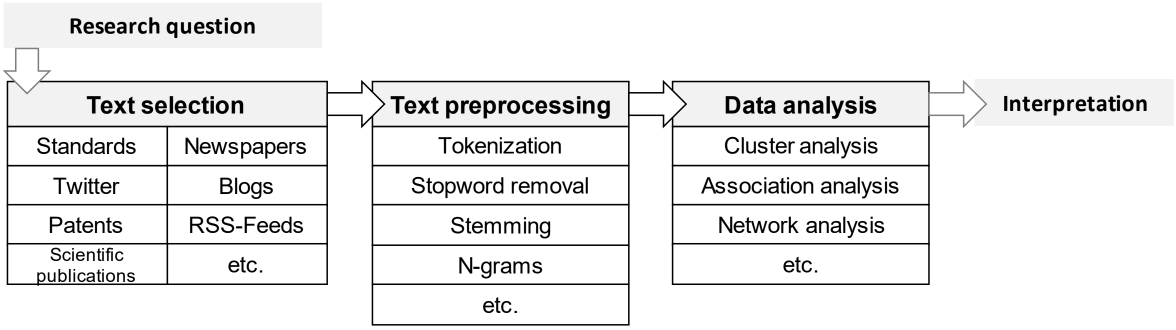


Figure 1-2 Text mining process

Text selection

A data source should be selected, which can answer the raised research question. The many possibilities range from social media to patents, standards and scientific publications (see following section for an overview). For formulating a precise search strategy, at least some principal knowledge of the subject or technology under consideration is necessary. The effort of this first step varies among data sources. While some data is retrieved from databases (e.g., patent, standards, scientific publications), other data requires manual gathering (e.g., reports).

Principally, each data source has its own strength and weaknesses. For example, measuring R&D activities requires that they are patented or published (e.g., Cunningham et al., 2006) or patents are submitted in different languages (Bonino et al., 2010).

Text preprocessing

Before text can be processed, it requires to be structured and transformed into a machine readable format. Therefore, the text is divided into its single elements as words (tokenization) and represented as a vector. To extract the relevant terms, mainly two different approaches are distinguished: working with stop words or the grammatical instances. Using with the grammatical instances, part of speech-tags are assigned to each word such as verb, article or noun. From this, relevant phrases or chains of words are extracted. Alternatively, stop words are used to remove irrelevant terms and function words (articles, conjunctions, pronouns, etc.). Then, further techniques as stemming (cuts each word to its basic form) or lemmatization (reduces word to root form based on dictionary) are applicable. Finally, independent of which strategy is used up to this point, the frequency of the terms is stored for further analysis (Manning et al., 2009).

Data analysis

For data analysis, in particular methods from statistics and data mining are applied such as classification and clustering (Han et al., 2012; Manning et al., 2009) and a wide range of software solutions exist (e.g., *R*, *RapidMiner*, *Weka*, *SPSS*, *Leximancer*). However, for a clear documentation of the research process, an own analysis software is more efficient. Then, the single process steps are traceable (see Kayser and Shala, 2014 for a further discussion). Therefore, a flexible framework that can be adapted to specific requirements, data sources and research questions is best.

Interpretation

Finally, interpreting the results is central, also because each dataset is subject to biases and limitations (e.g., completeness, representability). But data is not self-explanatory and cannot speak for themselves. Here, 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 they are intended for. Of course, text mining is an iterative process where the results raise further questions that require additional searches, additional data or follow-up research (e.g., interviews, workshops) to validate the results.

1.3 Using Text Mining for Foresight

This section will examine the contribution of text mining to foresight and describes different textual data sources. Then, it is argued how text mining might contribute to foresight and different applications are outlined.

1.3.1 Text as Data

Answering different research questions requires different data. This section introduces text sources that are or could be used in foresight.

Patents, scientific publications and standards

Patents, scientific publications and standards are used as indicator for technical change. Patent documents describe scientific and technical developments (e.g., Bonino et al., 2010). By definition, patents are state-of-the-art (in the moment they are published) and meanwhile protect technical solutions. Otherwise, scientific publications not only focus on technology but also include descriptions of basic and applied research considering a broader context. They also might describe ongoing work. As for standard documents (e.g., Goluchowicz and Blind, 2011), data is extracted from specialized databases that are quality assured and updated on a regular basis (e.g., Web of Science). With text mining, the abstracts and the full texts are accessible. These sources are frequently used in foresight to examine science and technology developments.

News articles

News articles inform society and contribute to public opinion making (e.g., Burkart, 2002). Their analysis may emphasize public concerns, beliefs and reservations. As for patents and scientific publications, the texts are edited and clearly written and therefore the same techniques for processing them are applicable. For example, Yoon (2012) examines web news for weak signals in the field of solar cells and appraises that, for his case, web news are a more refined and reliable source than blogs or web pages.

Social media

Social media is relevant for data gathering and participatory aspects. Principally, user-generated content such as blogs or Twitter may contribute insights from societal discourses. For example, Cachia et al. (2007) examine the potential of online social networks for foresight and trend recognition. They conclude that social networks indicate changes and trends in sentiment and social behavior and besides foster creativity and collective intelligence. Pang (2010) develops an approach to scan Web 2.0 contents produced by futurists on different web channels. *Social scanning* may deliver a very precise summary of what is discussed and what attracts *futurists* attention. Amanatidou et al. (2012) describe how they analyzed Twitter and other publicly available web sources with text mining in the context of weak signal identification and horizon scanning. Albert et al. (2015) analyze blogs with reference to technology maturity models and Glassey (2012) examines folksonomies, the tagging of web content with meta information, for their potential in early trend detection. Summarized, first applications exist based on different social media platforms for collecting information and user interaction. For text mining tasks, a wide range of applications to access and analyze this timely data evolve.

Websites in general

A lot of information is publicly available on websites. This data is semi-structured but might be analyzed by text mining. At the moment, a number of applications use web data related to innovation indicators. For example, company websites are retrieved to be examined about reports on innovations (Gök et al., 2015). Youtie et al. (2012) examine websites of small and medium enterprises in the field of nanotechnology regarding technology transition from discovery to commercialization. As will be introduced in Section 5, web mining can be applied in the context of scenario development. By aggregating the content of the websites, this form of data retrieval summarizes and describes the scenario field and serves as a starting point for discussing possible futures.

(Scientific) Reports and foresight studies

Foresight studies are a frequent information source in foresight exercises. Thus, they are manually screened for future statements. To automate this time-consuming task even partially, text mining would be of great value as for example Amanatidou et al. (2012) tried.

However, as they noticed, due to the length of the reports, the most frequent terms are not the most interesting, so cleaning and filtering tasks are necessary for weak signal detection. Kayser and Shala (2014) analyze reports with text mining to summarize the topic and deliver a starting point for the following scenario development.

1.3.2 Text Mining for Foresight Methods

Text mining as a quantitative approach might be a building block in foresight methods. The following describes existing (e.g., patent analysis) and possible future applications such as web-based scenario development.

Patent analysis

In recent years, there has been an growing interest in applying text mining methods for patent analysis to access the unstructured text fields as abstracts, claims or the descriptions (Masiakowski and Wang, 2013; Tseng et al., 2007; Abbas et al., 2014). As a main advantage compared to manual approaches, text mining aggregates large quantities of patents, generates further information as statistics or maps, and supports decision making (see, e.g., Wang et al., 2010). The current applications concentrate on various areas. For example, a number of work focuses on detecting patent infringements (Lee et al., 2013; Park et al., 2012). In addition, monitoring the R&D landscape is a common application. For example, Yoon et al. (2013) use patents to study the technology landscape and perform a competition analysis. Other applications are located in the context of technology transfer (Park et al., 2013b) or technology planning in general (Park et al., 2013a; Choi et al., 2012). Wang et al. (2010) design a framework to identify technology trends to guide R&D planning. TRIZ (Altshuller, 1984) supports the search for evolutionary patterns and a multi-step approach is applied. Automatic patent classification (as usually been performed manually) or its support is seen as a research trend (Bonino et al., 2010). For example, Cong and Loh (2010) propose a framework for rule-based patent classification. In response to the growing number of patent applications, Hido et al. (2012) try to automatically assess the quality of patent applications. Summarized, many research activities are conducted and patents are analyzed with text mining in different applications.

Publication analysis

Publication analysis examines scientific publications as the output of scientific work and measures developments and trends within science and technology. For decades, text mining is used in publication analysis (Cunningham et al., 2006; Kostoff, 2012). Text mining methods are applied on data fields as title, abstract, full text, keywords or for cleaning tasks but also on full texts. Different approaches for term extraction are applied as stop word 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). Methodologically, classification or cluster analysis (Glenisson et al., 2005; Delen and Crossland, 2008), topic modeling (e.g., Yau et al., 2014), or network and mapping approaches (e.g., van Eck and Waltman, 2011) are frequently applied. However, few studies compare datasets.

Technology roadmapping

Roadmaps are an instrument for strategic future planning (Barker and Smith, 1995; Möhrle et al., 2013). Related to roadmapping and text mining, some preliminary work 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 have been applied (e.g., SAO-based text analysis, text summarization, clustering) on different textual data sources (e.g., patents, product manuals). Together, these studies indicate that text mining and roadmapping are not conducted in parallel and text mining is merely done initially to get a thematic overview, but the core roadmapping is exclusively done by experts.

Fully integrated in roadmapping (see Figure 1-3), text mining and its results support each of the four process steps (Kayser et al., 2014). Continuous feedback loops between the two layers enrich the strategy process and serve as an objective base for balancing the internal views. In this framework, roadmapping is used for the internal strategy development and text mining for the analysis of external data and changes. Text mining supports the initial exploration and identification of relevant terms (step 1), detects trends on market and technology level (step 2) and indicates links between the objects of the roadmap (step 3).

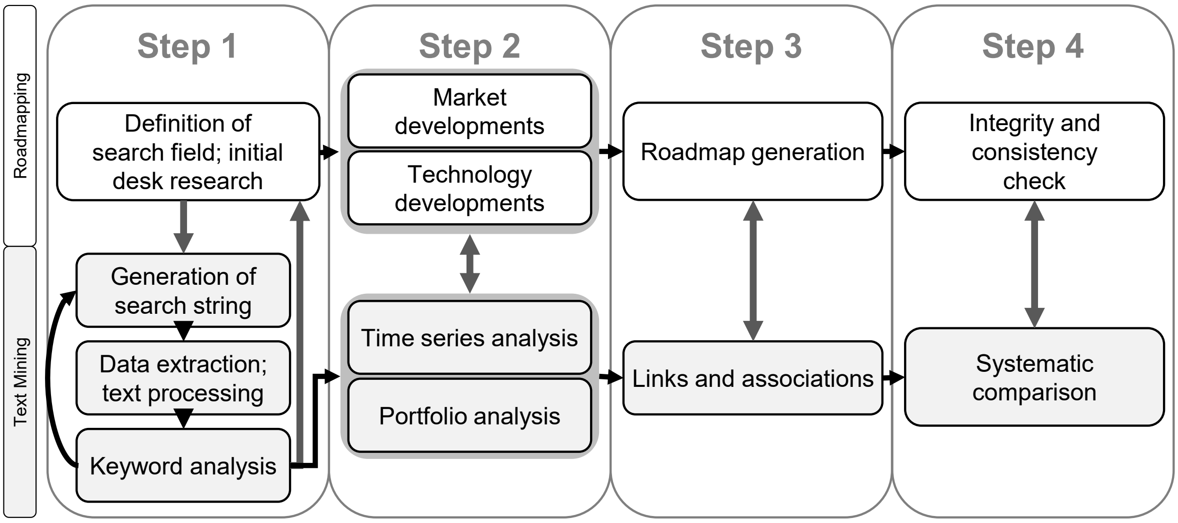


Figure 1-3 Process model: roadmapping and text mining (Kayser et al., 2014)

Figure 1-3 illustrates one way how to combine roadmapping and text mining. To provide a starting point and to illustrate the process, publication data was used. But in future work other data can be used - such as social media, reports, or news articles - to reinforce the customer perspective. By this process model, users can be integrated at different stages of the innovation process. For example, it is applicable for idea generation at the beginning of the innovation process or for a final alignment before the market entry.

Scenario development

Scenarios illustrate different futures each formulated as one scenario story. These stories serve as a framework to think about future challenges and developments influencing today's decisions (Reibnitz, 1991; van der Heijden, 2005). Among the many scenario approaches available at present, none uses text mining or seaks for more efficient ways to explore the scenario field, e.g., by automatic desk research. Principally, scenario development starts with desk research and literature analysis for a comprehensive understanding of the topic. Next, influence areas and future projections are formulated to describe the scenario field. These are combined to different scenario stories (second step) that are used for foresight in the third step.

Integrating text mining into the scenario preparation delivers a comprehensive overview on the topic and summarizes the scenario field very well (Kayser and Shala, 2014). This is illustrated in Figure 1-4. Applying text mining on the gathered literature facilitates structuring and organizing the scenario field. Practically, this reduces the reading effort and thereby the time effort for desk research and literature analysis at the beginning of the scenario process. Straight away discussions or workshops can start based on the results from text mining, e.g., to agree on influence factors and future projections.

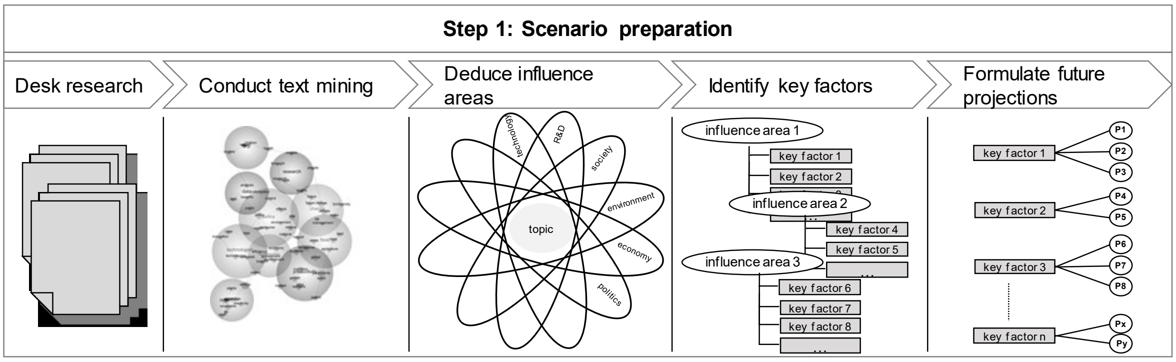


Figure 1-4 Scenario preparation including text mining (Kayser and Shala, 2014)

One of the advantages of this extended method are that more content and data can be analyzed than by classic literature analysis. Depending on the thematic scope, different data sources can be analyzed (e.g., reports, web mining, scientific publications). Finally, this enlarges the spectrum of foresight. In a more advanced case, desk research can be automated by retrieving content from platforms such as Twitter. For example, with web-based scenario development, more than 1.000 websites are processed - a number that is not to tackle manually (see Section 5).

1.4 Summed up: Relevancy of Text Mining for Foresight

This article describes the use of text mining for foresight and the contribution of different textual data sources. Up to now, text mining is, in particular, used for patent and publication analysis and less together with other foresight methods such as roadmapping or scenario development. Furthermore, opportunities evolve such as the usage of *new* data sources and other data analysis methods. The following section elaborates text mining for foresight and discusses the implications for foresight.

1.4.1 Assessing Text Mining for Foresight

Results of text mining might enable to reflect, check or validate intermediate results from the ongoing foresight activity. Potentially, text mining aids in better understanding ongoing changes and developments and their systemic implications. Reasons are the advanced data analysis and the larger number of accessible data sources. The following argues the contribution of text mining for foresight with reference to the process of text mining (see Section 1.2.2).

Access and text selection

As the previous literature overview showed, many applications combine text mining and patent- or publication analysis. Compared to that, few studies examine the potential of other data such as social media or the automatic analysis of scientific reports. So much textual data, such as news, social media*,* or classic websites, is not considered. By text mining, data sources are accessible such as Twitter (see Section 4) or web mining (see Section 5). Thereby, larger numbers of opinions might be integrated into foresight. This reduces the focus on science and technology und enables to address user aspects such as technology acceptance or concerns.

However, if *new* data is used in foresight, the strength and limitations should be clarified. So, for example, first the quality of Twitter as a data source needs to be examined and then can be used in the context of scenario development. For example, Twitter showed to be useful for retrieving data in real-time and for user engagement. It displays if there is a public debate, what is discussed and how. Principally, it enables the involvement of stakeholders not considered by foresight otherwise as well as rapid feedback on ideas. The automatic gathering of content with that variety and breadth is not possible with *classic* methods (e.g., interviews, workshops). Of course, Twitter data has limitations (e.g., representativity), but foresight requires diverse input and the results of Twitter analysis should be combined with other data and integrated in a larger foresight framework anyway.

Processing and structuring textual data

With text mining, data can be processed and structured that cannot be processed otherwise, particularly not in this volume and scale. This argument, for example, holds for news articles. Of course, news can be processed by content analysis (see for an overview on content analysis Krippendorff, 2013) and Twitter data can be manually gathered. However, this takes more time and smaller quantities can be processed. So manual and qualitative analysis encounter their limits and techniques as text mining are most relevant in our present time of increasing data volumes.

However, text mining cannot replace reading. Algorithms handle data different from reading and deliver a surface analysis. For example, it is not worth the effort to analyze Delphi statements with text mining because single statements about possible future developments cannot be analyzed in an automated manner. Therefore, some research questions still require qualitative and manual analysis.

Analyzing textual data

Text mining is applicable for comparing textual data as illustrated in Section 3 for news articles and scientific publications. Thereby, technology lifecycles and diffusion can be analyzed. Hypothesis about the evolution of a field are generated that should be proven by other methods.

1.4.2 Potential for Foresight

This article shows that different combinations of data sources and text mining approaches may contribute to foresight. For example, text mining helps to examine systemic links and the function of innovation systems or enhances the dynamic capabilities of firms (Kayser et al., 2014). The core advances are processing more content than without text mining and accessing data sources not used so far.

Moreover, the contributions of text mining can be argued on the example of the foresight process as described in Section 1.2.1. Principally, text mining is relevant for all three phases of foresight. As shown by the two methodological examples on roadmapping and future scenarios (that will be explained in more detail in the following sections), text mining can be a part of or used throughout the process. The three phases of foresight are also noticeable in the methods as summarized in Figure 1-5 and described in the following.

Input: Collecting and summarizing the available information is improved by text mining and more data can be processed. Thereby, more views and opinions are considered. Exploring and identifying relevant aspects in an objective manner is eased. Moreover, automated desk research reduces the time effort for summarizing the considered field and a greater scope can be captured (see Section 5 for details).

Process: One of the main contributions of text mining for 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. Particularly, for explorative foresight activities, this is valuable, because the process is built on a solid ground. In addition, results of text mining reflect, check or validate intermediate results from the ongoing foresight activity and thereby contribute to generating future knowledge. To get insights about possible future developments, text mining contributes by highlighting recent trends. Text mining may contribute an external perspective and serve for reflections. The results of text mining serve as a starting point for discussing possible futures promoting a creative discourse, in particular by hinting towards former disregarded aspects.

Output: For the final phase of foresight or even throughout the foresight process, the results of text mining are valuable to quantify and underline statements made. This aids in decision making and strategy planning.

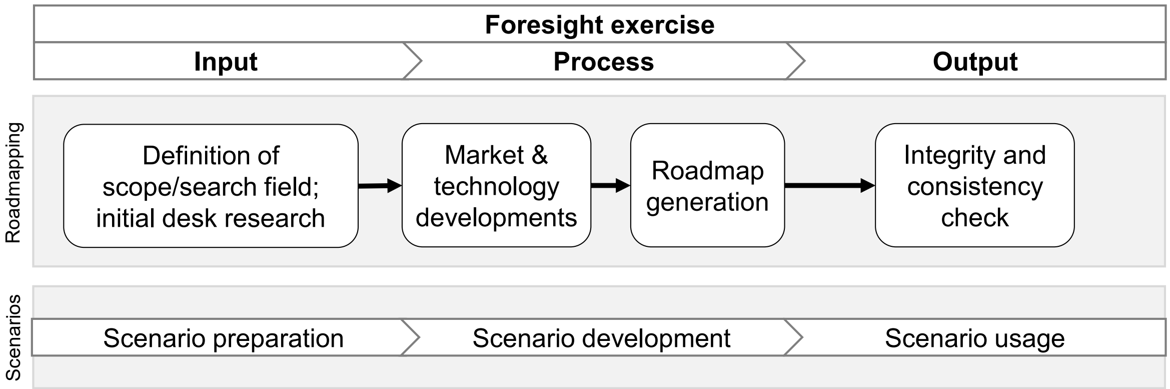


Figure 1-5 Foresight process model

This article describes the potential of combining foresight and text mining and outlines a framework for the following four articles where detailed examples are illustrated. These examples show how foresight can be extended and improved by foresight. Critical reflections will be part of the conclusion (Part 3). In this final part, limitations of text mining (and this thesis) are discussed and direction for future research are outlined.

References

Abbas, Assad; Zhang, Limin; Khan, Samee U. “A literature review on the state-of-the-art in patent analysis.” *World Patent Information* 37 (2014): 3–13.

Albert, Till; Möhrle, Martin G; Meyer, Stefan. “Technology maturity assessment based on blog analysis.” *Technological Forecasting and Social Change* 92 (2015): 196–209.

Altshuller, Genrich S. *Creativity as an exact science: The theory of the solution of inventive problems*. New York: Gordon and Breach Science Publishers, 1984.

Amanatidou, Effie; Butter, Maurits; Carabias, Vicente; Konnola, Totti, et al. “On concepts and methods in horizon scanning: Lessons from initiating policy dialogues on emerging issues.” *Science and Public Policy* 39, no. 2 (2012): 208–221.

Andersen, Allan D; Andersen, Per D. “Innovation system foresight.” *Technological Forecasting and Social Change* 88 (2014): 276–286.

Bañuls, Victor A; Salmeron, Jose L. “Scope and design issues in foresight support systems.” *International Journal of Foresight and Innovation Policy* 7, no. 4 (2011): 338–351.

Barker, Derek; Smith, David J. H. “Technology foresight using roadmaps.” *Long Range Planning* 28, no. 2 (1995): 21–28.

Bell, Wendell. *Foundations of futures studies: Human science for a new era*. New Brunswick, NJ [etc.]: Transaction Publishers, 2003.

Bonino, Dario; Ciaramella, Alberto; Corno, Fulvio. “Review of the state-of-the-art in patent information and forthcoming evolutions in intelligent patent informatics.” *World Patent Information* 32, no. 1 (2010): 30–38.

Bornmann, Lutz; Mutz, Rüdiger. “Growth rates of modern science: A bibliometric analysis.” *CoRR* abs/1402.4578 (2014).

Burkart, Roland. *Kommunikationswissenschaft: Grundlagen und Problemfelder*. 4th ed. Wien [et al.]: Böhlau, 2002.

Cachia, Romina; Compañó, Ramón; Da Costa, Olivier. “Grasping the potential of online social networks for foresight.” *Technological Forecasting and Social Change* 74, no. 8 (2007): 1179–1203.

Choi, Sungchul; Kim, Hongbin; Yoon, Janghyeok; Kim, Kwangsoo; Lee, Jae Y. “An SAO-based text-mining approach for technology roadmapping using patent information.” *R&D Management* 43, no. 1 (2013): 52–74.

Choi, Sungchul; Park, Hyunseok; Kang, Dongwoo; Lee, Jae Y; Kim, Kwangsoo. “An SAO-based text mining approach to building a technology tree for technology planning.” *Expert Systems with Applications* 39, no. 13 (2012): 11443–11455.

Cong, He; Loh, Han T. “Pattern-oriented associative rule-based patent classification.” *Expert Systems with Applications* 37, no. 3 (2010): 2395–2404.

Cuhls, Kerstin. “From forecasting to foresight processes—new participative foresight activities in Germany.” *Journal of Forecasting* 22, 2-3 (2003): 93–111.

Cuhls, Kerstin. “Schnittstellen von Foresight und Innovationsmanagement.” In *Zukunftsorientierung in der Betriebswirtschaftslehre*. 1st ed., edited by Victor Tiberius. Wiesbaden: Gabler, 2011: 189–199.

Cunningham, Scott W; Porter, Alan L; Newman, Nils C. “Special issue on tech mining: Tech Mining: Exploiting Science and Technology Information Resources.” *Technological Forecasting and Social Change* 73, no. 8 (2006): 915–922.

Da Costa, Olivier; Warnke, Philine; Cagnin, Cristiano; Scapolo, Fabiana. “The impact of foresight on policy-making: insights from the FORLEARN mutual learning process.” *Technology Analysis & Strategic Management* 20, no. 3 (2008): 369–387.

Delen, Dursun; Crossland, Martin D. “Seeding the survey and analysis of research literature with text mining.” *Expert Systems with Applications* 34, no. 3 (2008): 1707–1720.

van Eck, Nees J; Waltman, Ludo. “Text mining and visualization using VOSviewer.” *arXiv preprint arXiv:1109.2058 (*2011).

van Eck, Nees J; Waltman, Ludo; Noyons, Ed C. M; Buter, Reindert K. “Automatic term identification for bibliometric mapping.” *Scientometrics* 82, no. 3 (2010): 581–596.

Eerola, Annele; Miles, Ian. “Methods and tools contributing to FTA: A knowledge-based perspective.” *Futures* 43, no. 3 (2011): 265–278.

Eisenhardt, Kathleen M; Martin, Jeffrey A. “Dynamic capabilities: what are they?” *Strategic Management Journal* 21, 10-11 (2000): 1105–1121.

Feldman, Ronen; Sanger, James. *The text mining handbook: Advanced approaches in analyzing unstructured data*. Cambridge, New York: Cambridge University Press, 2008.

Fuller, Ted; Loogma, Krista. “Constructing futures: A social constructionist perspective on foresight methodology.” *Futures* 41, no. 2 (2009): 71–79.

Glassey, Olivier. “Folksonomies: Spontaneous crowd sourcing with online early detection potential?” *Futures* 44, no. 3 (2012): 257–264.

Glenisson, Patrick; Glänzel, Wolfgang; Janssens, Frizo; de Moor, Bart. “Combining full text and bibliometric information in mapping scientific disciplines.” *Information Processing & Management* 41, no. 6 (2005): 1548–1572.

Gök, Abdullah; Waterworth, Alec; Shapira, Philip. “Use of web mining in studying innovation.” *Scientometrics* 102, no. 1 (2015): 653–671.

Goluchowicz, Kerstin; Blind, Knut. “Identification of future fields of standardisation: An explorative application of the Delphi methodology.” *Technological Forecasting and Social Change* 78, no. 9 (2011): 1526–1541.

Han, Jiawei; Kamber, Micheline; Pei, Jian. *Data mining: Concepts and techniques*. 3rd ed. Amsterdam, Boston: Elsevier/Morgan Kaufmann, 2012.

Havas, Attila; Schartinger, Doris; Weber, Matthias. “The impact of foresight on innovation policy-making: recent experiences and future perspectives.” *Research Evaluation* 19, no. 2 (2010): 91–104.

van der Heijden, Kees. *Scenarios: The art of strategic conversation*. 2nd ed. Chichester, West Sussex, Hoboken, N.J.: John Wiley & Sons, 2005.

Hido, Shohei; Suzuki, Shoko; Nishiyama, Risa; Imamichi, Takashi, et al. “Modeling Patent Quality: A System for Large-scale Patentability Analysis using Text Mining.” *Information and Media Technologies* 7, no. 3 (2012): 1180–1191.

Hines, Andy; Gold, Jeff. “Professionalizing foresight: Why do it, where it stands, and what needs to be done.” *Journal of Futures Studies* 17, no. 4 (2013): 35–54.

Horton, Averil. “A simple guide to successful foresight.” *foresight* 1, no. 1 (1999): 5–9.

Huang, Lu; Zhang, Yi; Guo, Ying; Zhu, Donghua; Porter, Alan L. “Four dimensional Science and Technology planning: A new approach based on bibliometrics and technology roadmapping.” *Technological Forecasting and Social Change* 81, no. 0 (2014): 39–48.

Kayser, Victoria; Goluchowicz, Kerstin; Bierwisch, Antje. “Text Mininig for Technology Roadmapping: The strategic Value of Information.” *International Journal of Innovation Management* 18, no. 03 (2014): 1440004.

Kayser, Victoria; Shala, Erduana. “Generating Futures from Text: Scenario Development using Text Mining.” 5th International Conference on Future-Oriented Technology Analysis (FTA) - Engage today to shape tomorrow. Brussels, Belgium, 2014.

Kitchin, Rob. “Big Data, new epistemologies and paradigm shifts.” *Big Data & Society* 1, no. 1 (2014): 1–12.

Kostoff, Ronald N. “Text mining for science and technology - a review part I – characterization/scientometrics.” *Scientometrics* 1, no. 1 (2012): 11–21.

Krippendorff, Klaus. *Content analysis: An introduction to its methodology*. 3rd ed. Los Angeles, London: SAGE, 2013.

Lee, Changyong; Song, Bomi; Park, Yongtae. “How to assess patent infringement risks: a semantic patent claim analysis using dependency relationships.” *Technology Analysis & Strategic Management* 25, no. 1 (2013): 23–38.

Lee, Sungjoo; Lee, Seonghoon; Seol, Hyeonju; Park, Yongtae. “Using patent information for designing new product and technology: keyword based technology roadmapping.” *R&D Management* 38, no. 2 (2008): 169–188.

Linstone, Harold A. “Three eras of technology foresight.” *Technovation* 31, 2–3 (2011): 69–76.

Manning, Christopher D; Raghavan, Prabhakar; Schütze, Hinrich. *An Introduction to Information Retrieval*. New York: Cambridge University Press, 2009.

Martin, Ben R. “Foresight in science and technology.” *Technology Analysis & Strategic Management* 7, no. 2 (1995): 139–168.

Martin, Ben R; Johnston, Ron. “Technology Foresight for Wiring Up the National Innovation System.” *Technological Forecasting and Social Change* 60, no. 1 (1999): 37–54.

Martino, Joseph P. *Technological forecasting for decision making*. 3rd ed. New York: McGraw-Hill, 1993.

Masiakowski, Piotr; Wang, Sunny. “Integration of software tools in patent analysis.” *World Patent Information* 35, no. 2 (2013): 97–104.

Miner, Gary. *Practical text mining and statistical analysis for non-structured text data applications*. 1st ed. Waltham, MA: Academic Press, 2012.

de Miranda Santo, Marcio; Coelho, Gilda M; dos Santos, Dalci M; Filho, Lélio F. “Text mining as a valuable tool in foresight exercises: A study on nanotechnology.” *Technological Forecasting and Social Change* 73, no. 8 (2006): 1013–1027.

Möhrle, Martin G., Ralf Isenmann, and Robert Phaal, eds. *Technology roadmapping for strategy and innovation: Charting the route to success*. Berlin [et al.]: Springer, 2013.

Montoyo, Andrés; Martínez-Barco, Patricio; Balahur, Alexandra. “Subjectivity and sentiment analysis: An overview of the current state of the area and envisaged developments.” *Decision Support Systems* 53, no. 4 (2012): 675–679.

Öner, M. A. “On theory building in Foresight and Futures Studies: A discussion note.” *Futures* 42, no. 9 (2010): 1019–1030.

Ortner, Heike; Pfurtscheller, Daniel; Rizzolli, Michaela; Wiesinger, Andreas. “Zur Einführung – Datenflut und Informationskanäle.” In *Datenflut und Informationskanäle*. 1st ed., edited by Heike Ortner, Daniel Pfurtscheller, Michaela Rizzolli and Andreas Wiesinger. Innsbruck: Innsbruck Univ. Press, 2014.

Pang, Alex S.-K. “Social scanning: Improving futures through Web 2.0; or, finally a use for twitter.” *Global Mindset Change* 42, no. 10 (2010): 1222–1230.

Park, Hyunseok; Kim, Kwangsoo; Choi, Sungchul; Yoon, Janghyeok. “A patent intelligence system for strategic technology planning.” *Expert Systems with Applications* 40, no. 7 (2013a): 2373–2390.

Park, Hyunseok; Ree, Jason J; Kim, Kwangsoo. “Identification of promising patents for technology transfers using TRIZ evolution trends.” *Expert Systems with Applications* 40, no. 2 (2013b): 736–743.

Park, Hyunseok; Yoon, Janghyeok; Kim, Kwangsoo. “Identifying patent infringement using SAO based semantic technological similarities.” *Scientometrics* 90, no. 2 (2012): 515–529.

Popper, Rafael; Butter, Maurits. “How are foresight methods selected?” *foresight* 10, no. 6 (2008): 62–89.

Reibnitz, Ute. *Szenario-Technik: Instrumente für die unternehmerische und persönliche Erfolgsplanung*. Wiesbaden: Gabler, 1991.

Rohrbeck, René; Gemünden, Hans G. “Corporate foresight: Its three roles in enhancing the innovation capacity of a firm.” *Using Technological Intelligence for Strategic Decision Making in High Technology Environments* 78, no. 2 (2011): 231–243.

Salo, Ahti; Cuhls, Kerstin. “Technology foresight - past and future.” *Journal of Forecasting* 22, 2-3 (2003): 79–82.

Slaughter, Richard. *The foresight principle: Cultural recovery in the 21st century*. London, England: Adamantine Press; Adamantine, 1995.

Tidd, Joseph; Bessant, J. R. *Managing innovation: Integrating technological, market and organizational change*. 4th ed. Chichester, England, Hoboken, NJ: Wiley, 2009.

Tseng, Yuen-Hsien; Lin, Chi-Jen; Lin, Yu-I. “Text mining techniques for patent analysis.” *Information Processing & Management* 43, no. 5 (2007): 1216–1247.

Voros, Joseph. “A generic foresight process framework.” *foresight* 5, no. 3 (2003): 10–21.

Wang, Ming-Yeu; Chang, Dong-Shang; Kao, Chih-Hsi. “Identifying technology trends for R&D planning using TRIZ and text mining.” *R&D Management* 40, no. 5 (2010): 491–509.

Watts, Robert J; Porter, Alan L. “Innovation forecasting.” *Technological Forecasting and Social Change* 56, no. 1 (1997): 25–47.

Weiss, Sholom M. *Text Mining: Predictive methods for analyzing unstructured information*. New York: Springer, 2010.

Yau, Chyi-Kwei; Porter, Alan; Newman, Nils C; Suominen, Arho; Newman, Nils. “Clustering scientific documents with topic modeling.” *Scientometrics* 100, no. 3 (2014): 767–786.

Yoon, Byungun; Phaal, Robert; Probert, David R. “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, 2008: 417–422.

Yoon, Janghyeok. “Detecting weak signals for long-term business opportunities using text mining of web news.” *Expert Systems with Applications* 39, no. 16 (2012): 12543–12550.

Yoon, Janghyeok; Park, Hyunseok; Kim, Kwangsoo. “Identifying technological competition trends for R&D planning using dynamic patent maps: SAO-based content analysis.” *Scientometrics* 94, no. 1 (2013): 313–331.

Youtie, Jan; Hicks, Diana; Shapira, Philip; Horsley, Travis. “Pathways from discovery to commercialisation: using web sources to track small and medium-sized enterprise strategies in emerging nanotechnologies.” *Technology Analysis & Strategic Management* 24, no. 10 (2012): 981–995.

2 Text Mining for Technology Roadmapping: The strategic Value of Information

Abstract: Technology roadmapping is a well-established method used in strategy development and foresight. Furthermore, text mining offers untapped potentials concerning early detection and environmental scanning. Text mining can be used for structuring and analyzing thematic fields and facilitates an objective, quantitative summary of recent developments. In this article, the roadmapping process is split into different steps in order to analyze which text mining approaches might add further value. This leads to a two-layered process model and text mining is applied to systematically integrate external information in ongoing roadmapping processes. To demonstrate some of the benefits, the field of *cloud computing* is used to illustrate the procedure. As this article will show, the results extend the existing methodology, integrate an external view and complement expert opinion.

Keywords: Foresight, Roadmapping, Strategy Development, Text Mining, Publication Analysis, Association Analysis, Network Analysis

2.1 Introduction

The early detection of change and discontinuity in markets, society, and technology is essential for firms to remain globally competitive and is a core element of strategic foresight (Slaughter, 1997; Vecchiato and Roveda, 2010). One of the fundamental assumption of foresight is that the future is principally unknown, but today's decisions and anticipations can help to prepare and guide progress towards it (see e.g., Cuhls, 2003). Besides the detection of change and weak signals in the environment (Ansoff, 1975), adapting the strategic orientation to these changes is essential (e.g., Teece and Pisano, 1994; Eisenhardt and Martin, 2000). In this context, roadmapping is an established and frequently used method to adjust the strategy of a company by considering different future paths (see, e.g., Phaal et al., 2004).

However, one methodological challenge of roadmapping is the time-consuming initial exploration of the considered thematic field during the preparation phase. Further challenges are to relate strategic planning towards external developments, to guarantee that these developments are adequately considered in the strategy process and that the whole roadmapping process does not solely depend on input of (internal) experts and stakeholders. In addition, the volume of relevant internal and external information is constantly rising. To encounter these challenges, text mining offers unexploited potentials. Text mining is a structured method for the analysis of textual data (Feldman and Sanger, 2008; Manning et al., 2009). It might be used to explore and define the thematic scope, to locate and link roadmap objects, or to comprise additional sources of data. Consequently, text mining is expected to provide a base for comparison mirroring the ongoing roadmapping process with the results from a parallel data analysis. This might improve early trend detection and environmental scanning by recognizing ongoing changes and recent trends.

Therefore, the research question addressed in this article is how to combine technology roadmapping and text mining and to which extent roadmapping and its process steps are thereby assisted or even improved. In addition, the question of which text mining approaches should be applied to support the process steps best is examined. This finally leads to a two-layered process model with continuous feedback loops. The adapted process is explained on the example of *cloud computing.* For this case, scientific publication data is used in order to evaluate its contribution to technology roadmapping.

This article begins with an introduction of technology roadmapping and a summary of recent studies combining roadmapping and text mining (Section 2.2). Then, Section 2.3 describes the methodology. Finally, the results are discussed and conclusions are drawn in Section 2.4.

2.2 Thematic Background

The following describes the principles of technology roadmapping and introduces recent research combining text mining and roadmapping.

2.2.1 Technology Roadmapping

Foresight in general is an instrument that supports the induction of strategic decisions based on systematically looking into the longer-term future of a technology and its relations to science, economy and the environment and society (Martin, 1995). One of its main objectives is the identification of emerging technologies and related areas of strategic research plus taking a closer look at the benefits for the economy and society. Foresight with a focus on strategic planning contains activities evaluating future risks and challenges in order to improve the process of strategic decision-making. Dynamics in the entrepreneurial environment have to be recognized and suitably integrated in the strategy development of a firm (Slaughter, 1997). It is not only essential for firms to make decisions on the basis of internal sources and experts' opinions, but also to anticipate external information at an early stage. This enhances the dynamic capabilities of organizations defined as *"[...] ability to integrate, build, and reconfigure internal and external competences to address rapidly changing environments* (Teece et al., 1997)". This leads to a competitive advantage of the organization (Teece and Pisano, 1994; Eisenhardt and Martin, 2000). So strategic foresight further contributes by supporting this responsiveness and capability of firms. Looking ahead and analyzing current information can be used for decision-making or strategy development to prepare for future needs and opportunities (Coates et al., 2010; Vecchiato and Roveda, 2010).

Technology roadmapping has proven its use on different occasions but especially for strategy development to show alternative paths of technological futures (e.g., Barker and Smith, 1995; Phaal et al., 2004). For example, Möhrle and Isenmann (2013) define roadmapping as *"[...] graphical representation of technologies, often relating objects like products or competencies and the connections that have evolved between them in the course of time."* So, roadmaps are a planning tool used to target future objectives after reflecting and evaluating the alternative paths. Thereby, future technical developments and their interdependence with other aspects are examined. Given that, technology cannot be viewed isolated from external factors such as the market situation, internal capabilities, or the legal framework. Various formats exist to present and visualize roadmaps like tables, text or, most commonly, graphs with different layers, e.g., technology, market, and product (Phaal et al., 2004, 2010). In these graphs, timelines are generated and knowledge from different backgrounds is organized. Roadmaps weight up possible, likely and advantageous futures (Kappel, 2001) and create a consensus view of the future science and technology landscape by formulating strategies for decision makers (Kostoff and Schaller, 2001). In practice, roadmapping is applied on different topics, objectives, and focus (firm-specific, national, or international).

2.2.2 Text Mining for Technology Roadmapping

Text mining is a method to process and analyze textual data. It covers different algorithms merely based on linguistic and statistical methods to structure and analyze textual data (Weiss, 2010; Feldman and Sanger, 2008). The process can be summarized as follows. Initially, relevant textual data has to be identified such as abstracts of scientific articles, reports, or blog posts. This text is processed and transformed into a structured format. Following data analysis methods such as cluster analysis (Han et al., 2012) or association analysis (Agrawal et al., 1993) are applied to detect patterns and trends. The results are represented and visualized in an appropriate form, e.g., as term networks. Therefore, text mining is applicable to explore and define the thematic scope, to locate and link roadmap objects and thereby broadens the perspective by including additional sources of knowledge.

Summarized, text mining is an explorative approach for analyzing the state-of-the-art to derive statements about possible future developments. This is complemented by the character of roadmapping where strategic paths into the future are derived according to future visions and objectives. This might have a normative component. Thereby, the combination of text mining and roadmapping integrates explorative and normative thinking.

Roadmapping on the one hand and text mining techniques on the other hand are well-established and separately well-explored, but research of their inter-relation is still rare. However, several contributions exist that use text mining to support roadmapping which are introduced in the following. Evaluating the related literature showed that for exploring the thematic field, various text mining techniques are applied ranging from analyzing subject-action object (SAO) structures (Choi et al., 2013) to morphology analysis (Yoon et al., 2008a) or keyword-extraction techniques (Yoon et al., 2008b; Lin et al., 2008; Lee et al., 2008). For example, SAO structures originate from TRIZ (Altshuller, 1984), the Russian acronym for "*Theory of inventive problem solving*". SAO is focused on key concepts instead of keywords, where *AO* describes the problem description and S stands for the solution. As data source, patent data is used in most cases (e.g. Choi et al., 2013; Lee et al., 2008) but also other sources such as product manuals (Yoon et al., 2008a). For example, the patent data is used to analyze trends and identify relations between product and technology (Lee et al., 2008). Text mining is used preparatory to the roadmapping process to explore and structure the thematic field (Yoon et al., 2008b; Kostoff et al., 2004) or, for example, to develop a specific tech monitoring framework (Lin et al., 2008). In comparison, other results are technically far-reaching and the roadmap is constructed semi-automatically (Choi et al., 2013; Suh and Park, 2009). Even in these cases, dedicated expert involvement is still necessary, especially for controlling and selecting the keywords or search terms (Yoon et al., 2008a; Lee et al., 2008; Suh and Park, 2009). Text mining is rarely used in parallel to roadmapping as a simultaneous supporting element (Yoon et al., 2008a) so that the experts assist the data analysis (Lee et al., 2008) and not the other way round — data analysis supporting expert-based roadmapping.

In general, roadmapping often depends solely on expert statements and highly relies on the experts involved, their tacit knowledge and their participation in the process itself (Yoon et al., 2008b, 2008a). So the quality of the results depends on their professional knowledge or background (Specht and Mieke, 2004; Lee et al., 2008). Their role in roadmapping is inevitable for the final derivation of recommendations but also for interpreting text mining results. These people know the structures, workflows, and competencies as well as technical characteristics and dependencies between developments within technology, product and market within their own organizations. Nevertheless, the exclusive focus on this expertise for the development of the roadmap is challenging. Possibly, external developments in relevant areas might be missed out. As a further point, in the existing work, roadmapping is focused on exploring the thematic field but does not subdivide roadmapping in steps such as scope definition or linking objects to derive paths. This could be improved by differentiating the objectives of each step and considering how to enforce them by external input from text mining. Therefore, this article will show both, the value of analyzing external data with text mining and the contribution of text mining for each step of the roadmapping process.

2.3 Methodology

This section describes the methodology which systematically integrates text mining in the roadmapping process. This builds on the previous critique: The roadmapping process is not systematically split into phases to reflect the added value of special text mining techniques, but instead the process is considered as a whole. In this article, an attempt is made to resolve this observation by using the process model of Specht and Behrens (2005) as an orientation (see in Figure 2-1). This model has four steps. First, the process scope is defined and initial desk research is conducted. Second, trends on market and technology level are reflected. In the third step, the roadmap is generated and validated in the fourth step.

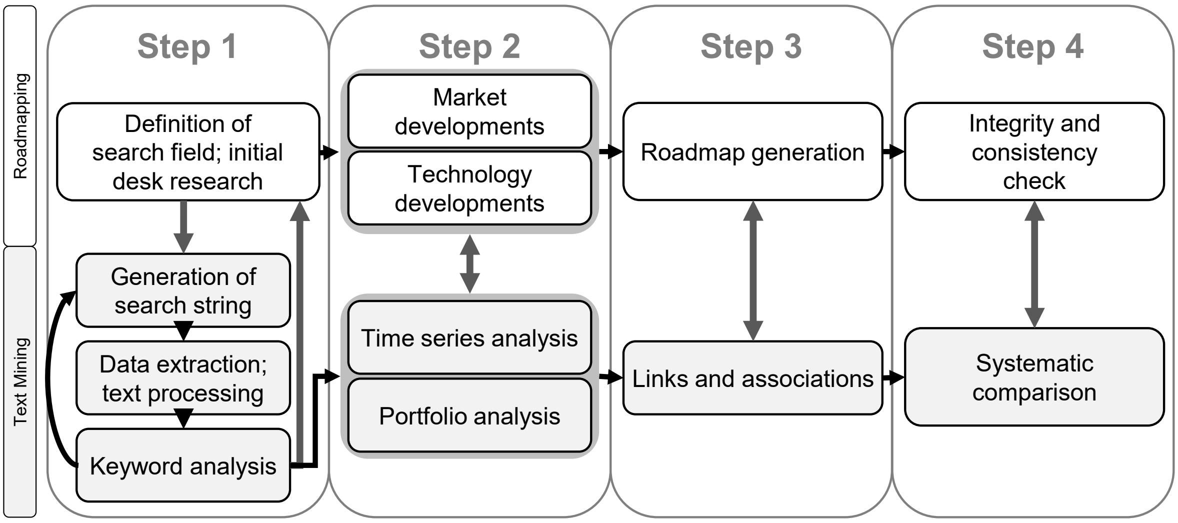


Figure 2-1 Process model (adaption of Specht and Behrens, 2005)

As indicated in Figure 2-1, a text mining layer is added to this model. Thereby, a combinatory approach is focused that considers internal expert knowledge together with external aspects in the form of results from the data analysis. This is enforced by continuous feedback loops between the two layers to ensure that the strategy development is in line with external developments because they are continuously reflected on internal considerations. The concrete implementation is explained in the following on the example of *cloud computing*.

2.3.1 Step 1: Scope of Roadmapping and first Examinations

In the first process step, the thematic scope of the roadmap and the considered layers (e.g., market, technology and product) are defined and a specific time horizon is fixed. Input from various sources can be used such as expert know-how, patents, studies, reports, or interviews for the clarification and delimitation of the scope and to identify first objects for the roadmap. As shown in Section 2.2.2, the previous work in this field mainly uses patent data, whereas, the text mining-process implemented in the following is based on data from scientific publications. Unlike patents, scientific publications also describe technological innovations that are not yet state-of-the-art and include further discussions, not merely technical details. Thereby, scientific publications cover a wider range of possible issues and bibliometric analysis provides thematic insights and indications of future developments (e.g., Grupp, 1997; Schmoch and Grupp, 1991). The data contains information for further analysis such as authors, title, year of publication, and particularly thematic information in the provided abstracts (e.g., Rooney et al., 2011; Blatt, 2009; van Eck and Waltman, 2007), subject categories (e.g., Leydesdorff et al., 2012), or keywords. The application of text mining for bibliometric analysis allows a profound analysis of data fields such as keywords or, in a more advanced application, the abstracts of the articles. Thus, text mining not only analyses the structured parts of the bibliographic data but especially processes the continuous text.

Data retrieval

For this article, besides the author keywords, the abstracts were analyzed for getting further input. However, the author keywords do not necessarily display the content of the article but rather the claims of the authors about the content (Delen and Crossland, 2008). To balance this issue, the abstracts were additionally processed. The abstracts contain a short summary of the main ideas and concepts described in the article and, thereby, give a more detailed view of the ongoing developments and discussions than the author keywords. Nevertheless, the text mining process is, in principle, adaptable to other data sources.

To retrieve the relevant data, a search strategy is generated. For improving the search strategy, the search results and the process scope should be compared leading to refinements of the scope and an iterative improvement of the search strategy. This validation has to be done by domain experts on the basis of the search results. In addition, the search results serve for a comparison to assess whether important aspects are missing or were overlooked in the process so far.

To illustrate the process, data to scientific articles related to *cloud computing* was retrieved from the *Web of Science*-database. This leads to 2.638 articles in total and covers the scientific output from 2007 to 2014. As described above, this article uses the abstracts and author keywords for further analysis. 2.586 articles featured an abstract and 2.171 articles were tagged by author keywords. In the following, abstracts and author keywords were separately processed.

Data processing

For the author keywords, a thesaurus was generated that summarizes synonyms, matches plural and singular forms, and replaces abbreviations or different spellings of terms such as American and British English. This thesaurus needs to be built manually. Thus, for example, *algorithm* and *algorithms* are merged or *virtualization* and *virtualisation*.

For processing the abstracts, an own Python-plugin was implemented. For each abstract, sentence wise the grammatical structure was examined to extract noun phrases (*Part of Speech*-extraction). This approach is more efficient than using tokenization and stemming on each word individually because many expressions are chains of words such as *quality of service* or *grid computing*. These are complex to extract with other text processing approaches. Therefore, regular expressions are formulated that filter out single nouns or chains of nouns (Bird et al., 2009). These regular expressions can be adapted to special linguistic requirements. For example, terms in the context of *cloud computing* have many "*as a*" constructs such as *software as a service* that are worth extraction. So chains of words matching these regular expressions are extracted from the text. In addition, plural forms and spellings can be cleaned automatically. Additionally, a stopword list removes common phrases with a low information gain such as *paper* or *article*. And the thesaurus (same as for the keywords) is applied for additional cleaning. This article used binary counting of terms per abstract.

Analyzing the results

To compare the abstract terms and author keywords, Figure 2-2 shows a word cloud for each data field. The word clouds map the 50 most frequent terms where the size of the term relates to the frequency of the other mapped terms (the more frequent the larger the term gets). For the following analysis, the term *cloud computing* was excluded because it occurred disproportionately often and would dominate the analysis otherwise.

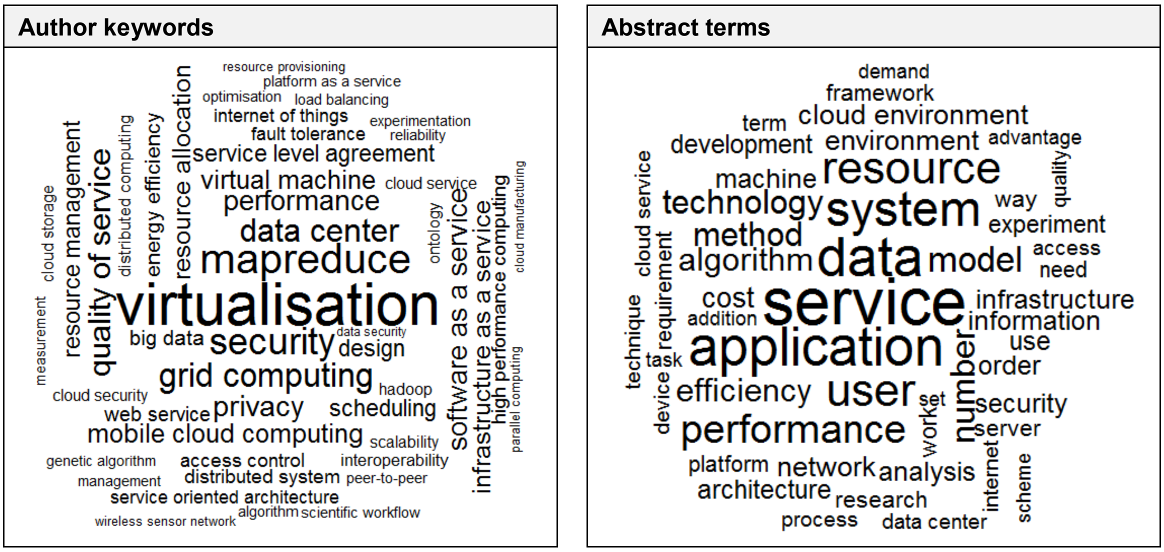


Figure 2-2 Word clouds for author keywords and abstract terms

Cloud computing is a new concept to offer scalable IT infrastructure on demand (e.g., Armbrust et al., 2010). This is technically realized by decentralized computing and storage based on virtualization. Nevertheless, *cloud computing* is not only a technological innovation but, for example, includes legal and privacy issues as well. This was also highlighted by the results in Figure 2-2. For example, among the 10 most frequent author keywords are *privacy* and *security*. Furthermore, concepts such as *virtualization*, *mapreduce* or *software as a service* are covered. The comparison shows that the author keywords are more specific than the abstract terms. This might explain their lower frequencies. For instance, the most frequent author keyword *virtualization* occurred 127 times while the most frequent abstract term *service* occurred 852 times. Among the most frequent terms from the abstracts are general terms such as *service*, *data*, *resource* and *performance*. However, *user* and *cost* are two aspects that are contained in the abstracts as well.

Implications for roadmapping

The information gained in this step supports the control and validation of the definition of the roadmapping scope. The results of the first step already indicate that the field of *cloud computing* is heterogeneous. This is relevant for the experts when deciding on the layers of the roadmap. Thus, not only technical aspects should be discussed when generating the roadmap but various other fields as well.

2.3.2 Step 2: Trends and Signals

The objective of the second step is to identify recent trends and developments in market and technology that might be of strategic and future relevance and therefore should be considered in the roadmapping process. To support this examination, timelines of related terms were retrieved and a portfolio analysis was conducted hinting towards recent trends and emerging technologies.

Timelines

To illustrate the growth of *cloud computing* with respect to related terms and technologies, timelines map some of the most frequent terms from the previous step. These are *grid computing*, *web service*, *virtualization, software as a service*, and *service-oriented architecture*. For these terms, own search strings were generated and related data was extracted from *Web of Science* for the time interval 2000-2014. The six searches were combined using the logical OR-operator that prevents double-counting of articles. As Figure 2-3 indicates, the OR-combination first co-evolved with *web service* and then with *cloud computing*.

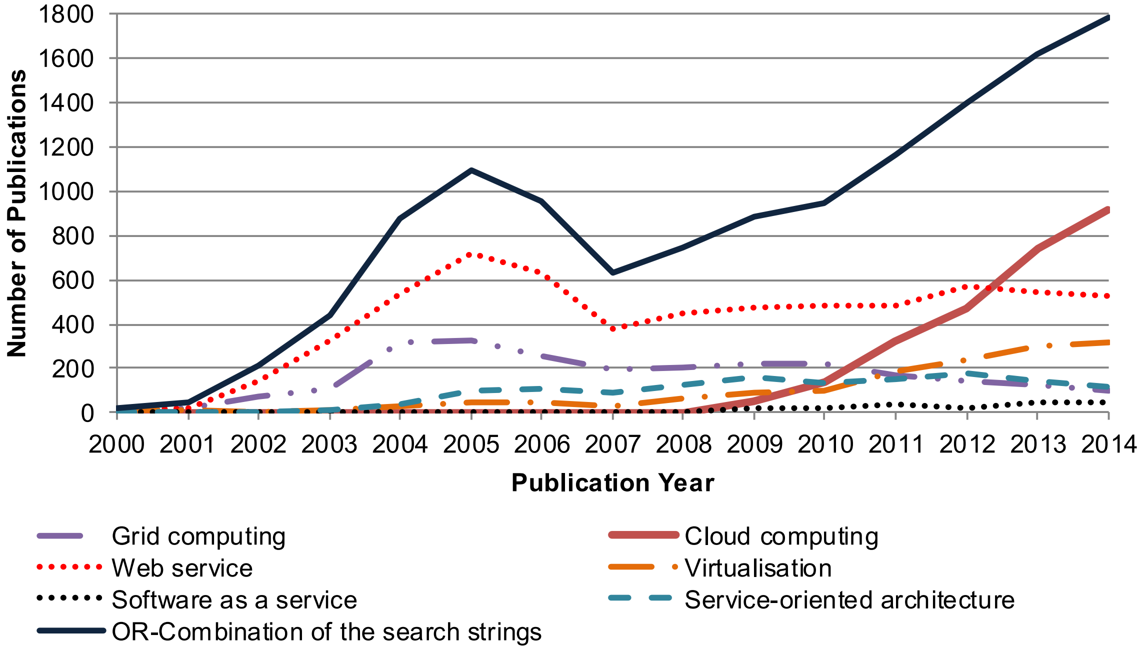


Figure 2-3 Timelines of related terms and technologies (source: Web of Science; time interval: 2000-2014)

As Figure 2-3 highlights, *cloud computing* is still a very young topic that originates in the considered data from the year 2007 onwards, while *grid computing*, *virtualization, web service* or *service-oriented architecture* continuously developed earlier.

Trend portfolio analysis

As a next step, a portfolio analysis delivers an overview on emerging and declining terms and helps to detect and assess trends and current developments (see, e.g., Choi et al., 2013). For this analysis, the considered time interval ranges from 2011 to 2014. This interval is divided in two time slices, *T (2011-2012)* and *T-1 (2013–2014)*. Next, two measures were taken into account to build the dimensions of the portfolio. First, the relative frequency of the terms is mapped on the x-axis. It is calculated by the frequency in the time interval *T* as $\sum Freq(K,T)$ and divided by the total number of articles *A* in the time interval *T* as $\sum Freq(A,T)$. Second, the growth of the terms is approximated and mapped on the y-axis. It is calculated by the distance of the term frequencies in the two time intervals divided by the frequency in the first interval. This means:

$$Approximated\ Growth\ (K) = \frac{\sum Freq(K,T) - \sum Freq(K,T-1)}{\sum Freq(K,T)}$$

The portfolios were separately calculated for the author keywords and the terms from the abstracts. The portfolio graphically subdivides the considered terms into four groups: *emerging* (upper left corner), *core* (upper right corner), *declining* (lower left corner) and *established* (lower right corner).

To begin with, Figure 2-4 maps the portfolio for selected author keywords. Implicitly, the four groups as described before are noticeable here. *Grid computing* as predecessor of *cloud computing* and for a long established technology is declining together with terms such as *software as a service* or *service oriented architecture*. In contrast, *internet of things* is among the emerging terms together with *big data* or *attribute-based encryption*. *Mobile cloud computing* already moves on to the core themes such as *mapreduce and virtualization,* the established terms. These few *established* terms might be due to the novelty of *cloud computing* and its emergence from 2007 onwards.

Next, Figure 2-5 illustrates the portfolio for the abstract terms. A set of basic terms such as *system*, *service* and *resource* frequently occur in the dataset and have been continuously mentioned over the last years and their growth rate is almost 0. The usage of the term *data* is still growing, may be due to the trend around big data and data-intense applications. Among the emerging terms is, for example, *device*. This might relate to the trend around mobile applications. Opposed to the first portfolio, in this portfolio *virtualization* is a declining term but has a negative growth rate in both cases.

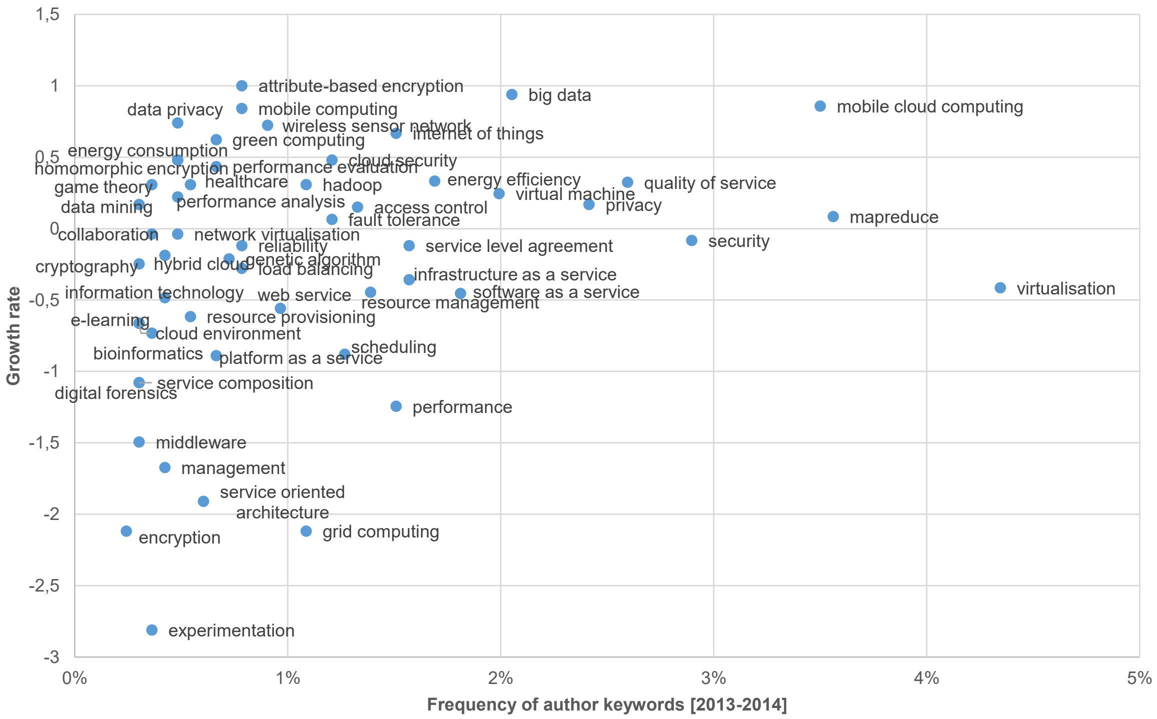


Figure 2-4 Trend portfolio of selected author keywords (excerpt)

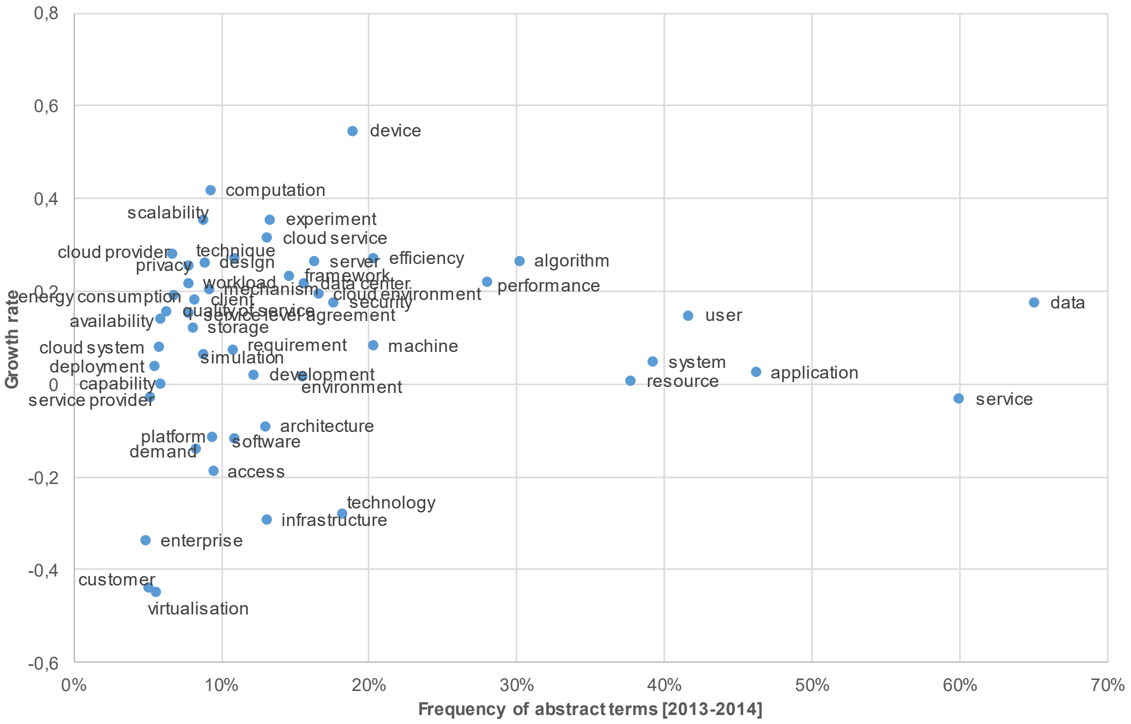


Figure 2-5 Trend portfolio of selected abstract terms (excerpt)

Implications for technology roadmapping

The portfolios and the timelines provide orientation about recent trends such as data-intense applications or internet of things. This might encourage a discussion about emerging technologies and serve as a basis for an objective comparison of statements made during the roadmapping process.

2.3.3 Step 3: Roadmap Generation

In the third step, the roadmap is generated. The objects are chronologically ordered and plausible development paths are derived. This step might be assisted by detecting groups and dependencies between the objects. In this particular case, association analysis and network analysis are conducted.

Association analysis

Association analysis is used to identify terms that frequently occur together in abstracts respectively keyword lists and identifies dependencies between terms in a dataset (Agrawal et al., 1993; Lopes et al., 2007). This supports deriving information about generalized or specialized relations because the relations are directed. For the analysis, frequent itemsets are constructed that indicate which terms often occur together. The two key measures are *support* and *confidence* for which thresholds are set to restrict the algorithm. The support is the relative frequency in which a term occurs in a dataset. The confidence of a rule ($A \rightarrow B$) describes the support $Sup(A \cup B)$ divided by the number of transactions including only A what means divided by $Sup(A)$:

$$conf(A \rightarrow B) = \frac{Sup(A \cup B)}{Sup(A)}$$

One popular example of this method is the market basket analysis which tries to identify how purchased items are associated (Han et al., 2012). In the case of publication data, each article is a transaction while the rules are derived for sets of terms. For this analysis, articles containing no terms were excluded.

The association rules for the author keywords are illustrated in Table 2-1. They are restricted by a confidence of 0.3 and a support of 0.005 which resulted in 13 rules. The relatively low support is due to the many articles and high number of different keywords. For example, *platform as a service* occurs together with *software as a service* in 57.7% of the cases when *platform as a service* occurs. Another example is the association between algorithm and performance (Confidence: 55%). This implies that when algorithm is mentioned performance is mentioned in addition and there is a dependency of the term *algorithm* on the term *performance*.

No.	Rule	Confidence
1	design ---> performance	0,667
2	performance ---> design	0,531
3	design ---> management	0,333
4	management ---> design	0,722
5	management ---> performance	0,611
6	measurement ---> design	0,647
7	design ---> virtualisation	0,436
8	algorithm ---> design	0,650
9	design ---> algorithm	0,333
10	platform as a service ---> software as a service	0,579
11	algorithm ---> performance	0,550
12	design ---> performance, virtualisation	0,333
13	design ---> performance, virtualisation:	0,333

Table 2-1 Association rules for author keywords [conf: 0.3; sup: 0.005]

The association analysis for the abstract terms with a confidence of 0.3 and a support of 0.05 delivered 19 rules (see Table 2-2). The support value is lower for this case because the higher frequency of the abstract terms imply that they are contained in more transactions. The rule *quality* => *service* means that they with a confidence of 78,5% occur together when quality is contained. This implies a reliance of the term *quality* on *service*. Further on, for *service* and *application* (rule 8 and 9), there is not much difference in the confidence value. In comparison, *data* is frequently mentioned together with *user*, *system* or *performance* (rule 3, 4, and 6).

No.	Rule	Confidence
1	system ---> application	0,302
2	resource ---> service	0,430
3	user ---> data	0,341
4	system ---> data	0,318
5	user ---> application	0,303
6	performance ---> data	0,306
7	system ---> service	0,330
8	service ---> application	0,310
9	application ---> service	0,377
10	model ---> service	0,385
11	performance ---> application	0,329
12	user ---> service	0,451
13	performance ---> system	0,306
14	infrastructure ---> service	0,554
15	technology ---> service	0,363
16	quality ---> service	0,785
17	resource ---> application	0,351
18	performance ---> service	0,310
19	number ---> service	0,364

Table 2-2 Association rules for abstract terms [conf: 0.3; sup: 0.05]

Term networks

Next, the co-occurrence of terms is visualized in a term network (Bastian et al., 2009). Here, the nodes represent terms and the edges indicate a relation between two considered terms. The node size depends on the node degree as the number of edges a node has (Wasserman and Faust, 2007). While the number of association rules is restricted by the support and confidence value, filters related to the node degree or node frequency are applied in order to reduce the size of the network graph. The algorithm *force atlas* was used to structure the graph. To handle the complexity, the term networks are built for the 100 most frequent terms per dataset.

Figure 2-6 shows the term network for author keywords. *Grid computing* is a well-connected term in the network with frequent links to *virtualization* or *software as a service*. *Quality of service* has a relatively central position in the network and is tightly linked to *service level agreements*. The link between *security* and *privacy* indicates that these issues are frequently addressed together in this context.

The network of the abstract terms in Figure 2-7 contains a highly connected set of basic cloud computing vocabulary such as *service*, *system* and *application.* As already indicated by Figure 2-2, the abstract terms frequently contain more general terms such as *algorithm, architecture,* or *infrastructure*.

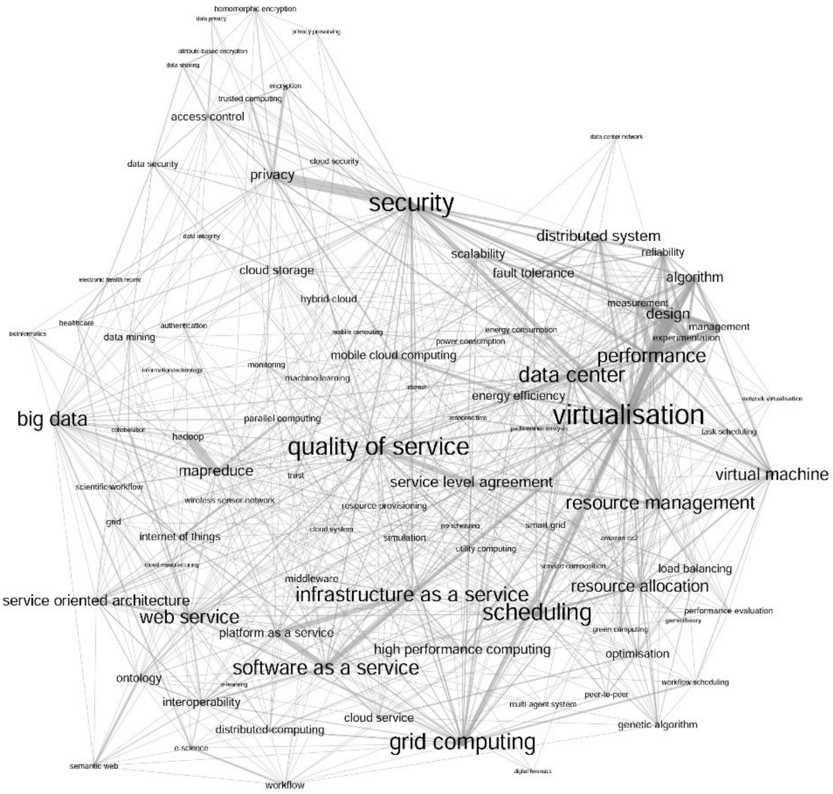


Figure 2-6 Term network for author keywords (node degree ≥ 4)

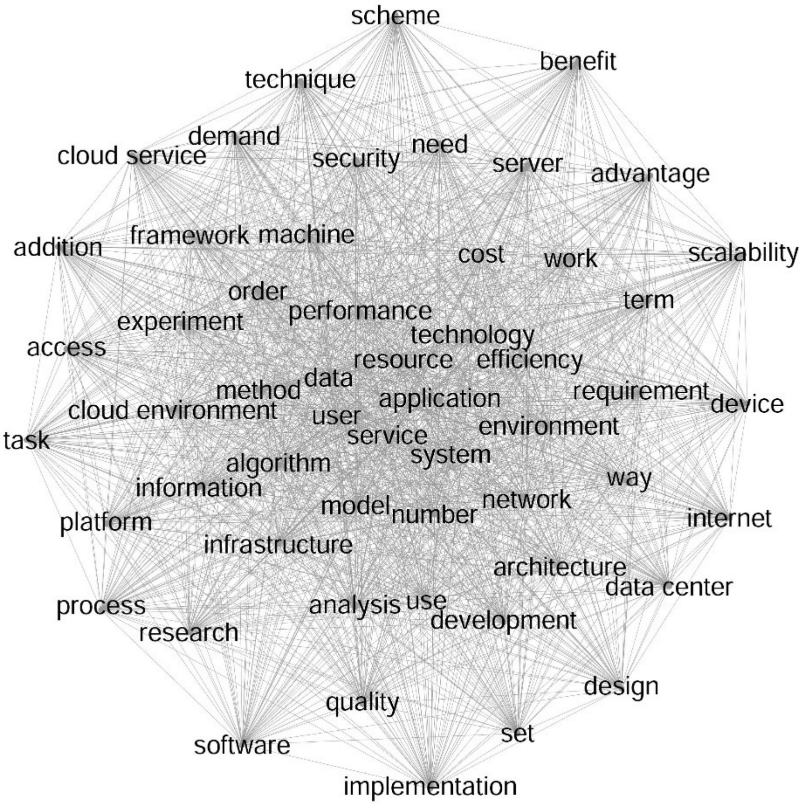


Figure 2-7 Term network for abstract terms (node frequency ≥ 170)

The comparison of the two networks shows that the abstract terms seem to be of a higher degree and are stronger interconnected than the author keywords. One explanation is that abstracts are longer text fragments than the author keywords which have fewer entries per article. For future work, networks from full texts should not rely on the degree as node size and other measures should be used at this point.

Implications for technology roadmapping

The networks indicate links between roadmap layers as, for example, between *security* as a legal or societal aspect and *virtualization* on the technical layer. Therefore, the information generated in this third step assists in ordering and linking the objects while an exact placement on the roadmap is not achieved by the analysis conducted here. The position of the objects on the time scale is also a question of strategic interests and decisions and depends on internal considerations. Thus, the exact placement of the identified objects on the roadmap still has to be discussed with the domain experts. Nevertheless, additional information about connections between the objects was obtained by these analyses.

To further illustrate the method designed in this article, Figure 2-8 maps an exemplary roadmap path. For example, a company wants to meet the growing demand for large scale applications and therefore offers a cloud service. *Hadoop* as one implementation offers the technical requirements as software solution with *mapreduce* as technical basis. According to the trend portfolio in step 2, *hadoop* is newly evolving and step 3 showed its relation to *mapreduce*, a frequent term in the *cloud computing*-context. This information might guide the experts in their planning process. As a final consequence, new business applications might evolve and the market share might increase. This very broad case shows that the analysis of step 1 to 3 prove that the future planning is in line with the developments in the outside and adequately considers the latest developments.

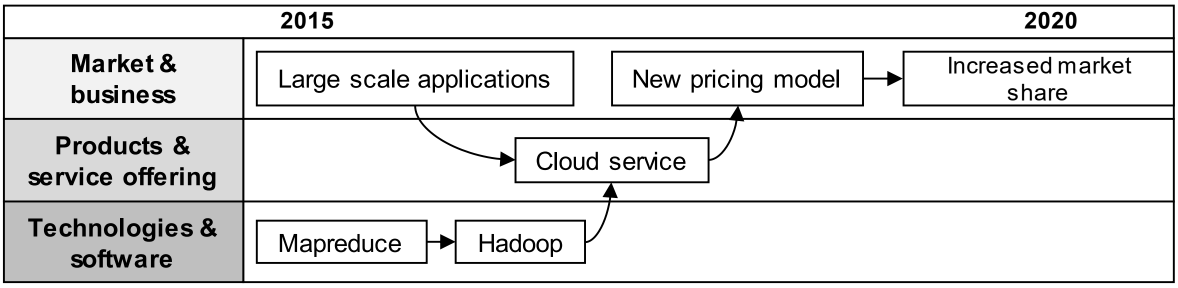


Figure 2-8 Exemplary path of a *cloud computing*-roadmap

2.3.4 Step 4: Roadmap Validation and Consistency Check

Finally, the identified relations and dependencies between and within the layers are checked. Therefore, the consistency and level of completeness of the generated roadmap is examining (e.g., plausibility of the links, chronological course). Strategic measures are derived to conclude the process. In this step, no explicit text mining analysis is conducted, but the results from the three previous steps are taken for final adjustments. The main questions for this final validation refer, for example, to the level of preciseness (e.g., degree of detail; missing aspects) and the plausibility of the detected paths as well as potential gaps and inconsistencies.

2.4 Discussion and Conclusion

In this article, a new process was developed for combining roadmapping and text mining. The objective of this article was to develop a process model that parallelizes technology roadmapping and text mining in order to consider external changes in internal strategy processes. Text mining was used to access external data, while roadmapping was used for strategy formulation and implementation. In contrast to the related literature, this approach divides roadmapping in four steps and applies adequate text mining methods within each. Text mining in this context supports the initial exploration and identification of terms of interest as visualized in the first step, detects trends as in the second step and offers methods to derive paths in the third step. In particular, text mining structures the considered thematic field and balances experts' statements.

In more detail, abstracts and author keywords were used in this article to gain thematic insights. Generally, the initial effort for processing text differs. For both data fields, a thesaurus needs to be build that matches synonyms and variations. For the author keywords, it is sufficient to apply a thesaurus, but singular and plural forms need to be merged manually. Processing abstracts for the first time brings a high effort because of the time needed for programming, but then this process can be automated. For example, plural forms can be cleaned in an automatic manner. In the case of *cloud computing*, the comparison of both data fields shows that abstracts include more basic descriptive terms, while the author keywords have a wider thematic focus by containing terms as *security* or *privacy.* However, term networks show better results for the keywords because of the lower density. In addition, the trend portfolio for the author keywords seems to be more responsive and contains less common words. Therefore, for each dataset it should be tested and decided which data to use in the specific case and other topics should be investigated. Nevertheless, processing abstracts is an important preparatory step for getting experience in analyzing longer texts such as reports in future work.

As a further point, the used text processing requires further improvement and should be seen as a starting point rather than a finished method. This applies for example to cleaning and synonym detection (Manning et al., 2009) or statistical measures that evaluate the importance of terms (Berry, 2004). In general, the complexity of text mining highly depends on how structured the input data is. Bibliometric data is already very structured, but processing the content of whole reports is of higher complexity due to, for example, figures and structuring the input.

Text mining can be used to access a wide range of data sources, but their application and quality in roadmapping, especially their focus and restrictions are partly unknown. Publication data as a frequent source in strategic foresight (see, e.g., Mietzner and Reger, 2009) was used in this article to illustrate the process, but this data displays the science perspective. Aspects as public or political discourse and explicit user demand are underrepresented. Other common sources in strategic foresight such as specialized databases or archives (see, e.g., Müller and Müller-Stewens, 2009) should be examined in future work. Furthermore, newspaper articles, Twitter, or blogs are an option to emphasize the demand side and user aspects. A next step would be the strategic exploitation of these alternatives. This is also most relevant for foresight and strategic planning to meet the increasing volume of textual data and make more accurate and informed statements about the future. In general, the strategic decision process concerning future investments or efforts in research and development activities, services, and products is supported by this objective and quantitative text analysis.

Developing a roadmap in a completely automatic manner is not possible but rather a hybrid approach to assist expert-based roadmapping (see also Yoon et al., 2008a; Kostoff and Schaller, 2001). Otherwise, this would lead to uniform solutions that disregard individual aspects and interests of the unit conducting roadmapping. The method developed in this article assists roadmapping by indicating related terms and giving an overview on ongoing developments and trends. The text mining results serve as objective supplement to the experts' statements made during the process. Certainly, different ways of integrating text mining in the roadmapping process exist. In the end, this model is one possible way to combine roadmapping and text mining, leaving the exact implementation to the user. Thus, the extent to which the text mining results are used for roadmapping will vary from case to case. Nevertheless, this model provides opportunities to add an external view and, thereby, orients the strategy development on external developments. This strengthens early detection and possibly enhances dynamic capabilities.

References

Agrawal, Rakesh; Imieliński, Tomasz; Swami, Arun. "Mining association rules between sets of items in large databases." Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data. Washington, D.C.,USA, 1993: 207–216.

Altshuller, Genrich S. *Creativity as an exact science: The theory of the solution of inventive problems*. New York: Gordon and Breach Science Publishers, 1984.

Ansoff, H. I. "Managing strategic surprise by response to weak signals." *California management review* 18, no. 2 (1975): 21–33.

Armbrust, Michael; Fox, Armando; Griffith, Rean; Joseph, Anthony D., et al. "A view of cloud computing." *Communications of the ACM* 53, no. 4 (2010): 50–58.

Barker, Derek; Smith, David J. H. "Technology foresight using roadmaps." *Long Range Planning* 28, no. 2 (1995): 21–28.

Bastian, Mathieu; Heymann, Sebastien; Jacomy, Mathieu. "Gephi: An Open Source Software for Exploring and Manipulating Networks." Conference: Proceedings of the Third International Conference on Weblogs and Social Media. San Jose, California, USA, 2009.

Berry, Michael W. *Survey of text mining: Clustering, classification, and retrieval*. New York: Springer, 2004.

Bird, Steven; Klein, Ewan; Loper, Edward. *Natural language processing with Python*. 1st ed. Beijing, Cambridge [Mass.]: O'Reilly, 2009.

Blatt, Eli M. "Differentiating, describing, and visualizing scientific space: A novel approach to the analysis of published scientific abstracts." *Scientometrics* 80, no. 2 (2009): 385–406.

Choi, Sungchul; Kim, Hongbin; Yoon, Janghyeok; Kim, Kwangsoo; Lee, Jae Y. "An SAO-based text-mining approach for technology roadmapping using patent information." *R&D Management* 43, no. 1 (2013): 52–74.

Coates, Joseph; Durance, Philippe; Godet, Michel. "Strategic Foresight Issue: Introduction: Strategic Foresight." *Technological Forecasting and Social Change* 77, no. 9 (2010): 1423–1425.

Cuhls, Kerstin. "From forecasting to foresight processes—new participative foresight activities in Germany." *Journal of Forecasting* 22, 2-3 (2003): 93–111.

Delen, Dursun; Crossland, Martin D. "Seeding the survey and analysis of research literature with text mining." *Expert Systems with Applications* 34, no. 3 (2008): 1707–1720.

van Eck, Nees J; Waltman, Ludo. "Bibliometric mapping of the computational intelligence field." *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 15, no. 05 (2007): 625–645.

Eisenhardt, Kathleen M; Martin, Jeffrey A. "Dynamic capabilities: what are they?" *Strategic Management Journal* 21, 10-11 (2000): 1105–1121.

Feldman, Ronen; Sanger, James. *The text mining handbook: Advanced approaches in analyzing unstructured data*. Cambridge, New York: Cambridge University Press, 2008.

Grupp, Hariolf. *Messung und Erklärung des technischen Wandels: Grundzüge einer empirischen Innovationsökonomik*. Berlin [et al.]: Springer, 1997.

Han, Jiawei; Kamber, Micheline; Pei, Jian. *Data mining: Concepts and techniques*. 3rd ed. Amsterdam, Boston: Elsevier/Morgan Kaufmann, 2012.

Kappel, Thomas A. "Perspectives on roadmaps: how organizations talk about the future." *Journal of Product Innovation Management* 18, no. 1 (2001): 39–50.

Kostoff, Ronald N; Boylan, Robert; Simons, Gene R. "Disruptive technology roadmaps." *Roadmapping: From Sustainable to Disruptive Technologies* 71, 1–2 (2004): 141–159.

Kostoff, Ronald N; Schaller, Robert R. "Science and technology roadmaps." *Engineering Management, IEEE Transactions on* 48, no. 2 (2001): 132–143.

Lee, Sungjoo; Lee, Seonghoon; Seol, Hyeonju; Park, Yongtae. "Using patent information for designing new product and technology: keyword based technology roadmapping." *R&D Management* 38, no. 2 (2008): 169–188.

Leydesdorff, Loet; Rotolo, Daniele; Rafols, Ismael. "Bibliometric Perspectives on Medical Innovation using the Medical Subject Headings (MeSH) of PubMed." *CoRR* abs/1203.1006 (2012).

Lin, F; Wei, C; Lin, Y; Shyu, Y. "Deriving Technology Roadmaps with Tech Mining Techniques." Proceedings of the PACIS, 2008.

Lopes, A. A; Pinho, R; Paulovich, F. V; Minghim, R. "Visual text mining using association rules." *Computers & Graphics* 31, no. 3 (2007): 316–326.

Manning, Christopher D; Raghavan, Prabhakar; Schütze, Hinrich. *An Introduction to Information Retrieval*. New York: Cambridge University Press, 2009.

Martin, Ben R. "Foresight in science and technology." *Technology Analysis & Strategic Management* 7, no. 2 (1995): 139–168.

Mietzner, Dana; Reger, Guido. "Practices of Strategic Foresight in Biotech Companies." *International Journal of Innovation Management* 13, no. 02 (2009): 273–294.

Möhrle, Martin G; Isenmann, Ralf. "Basics of Technology Roadmapping." In *Technology roadmapping for strategy and innovation: Charting the route to success,* edited by Martin G. Möhrle, Ralf Isenmann and Robert Phaal. Berlin [et al.]: Springer, 2013: 1–10.

Müller, Adrian W; Müller-Stewens, Günter. *Strategic Foresight: Trend- und Zukunftsforschung in Unternehmen - Instrumente, Prozesse, Fallstudien*. Stuttgart: Schaeffer Poeschel; Schäffer-Poeschel Verlag für Wirtschaft Steuern Recht GmbH, 2009.

Phaal, Robert; Farrukh, Clare; Probert, David R. "Technology roadmapping—A planning framework for evolution and revolution." *Roadmapping: From Sustainable to Disruptive Technologies* 71, 1–2 (2004): 5–26.

Phaal, Robert; Farrukh, Clare; Probert, David R. *Roadmapping for strategy and innovation: Aligning technology and markets in a dynamic world*. Cambridge: University of Cambridge, Institute for Manufacturing, 2010.

Rooney, David; McKenna, Bernard; Barker, James R; Rooney, D., et al. "History of Ideas in Management Communication Quarterly." *Management communication quarterly* 25, no. 4 (2011): 583–611.

Schmoch, Ulrich; Grupp, Hariolf. "Technologieindikatoren: Aussagekraft, Verwendungsmöglichkeiten, Erhebungsverfahren." In *Handbuch des Informationsmanagements im Unternehmen: Technik, Organisation, Recht, Perspektiven,* edited by Hans-Jörg Bullinger. München: Beck, 1991: 1571–1615.

Slaughter, Richard A. "Developing and applying strategic foresight." *ABN Report* 5, no. 10 (1997): 13–27.

Specht, Dieter; Behrens, Stefan. "Strategische Planung mit Roadmaps — Möglichkeiten für das Innovationsmanagement und die Personalbedarfsplanung." In *Technologie-Roadmapping: Zukunftsstrategien fur Technologieunternehmen*. 2nd ed., edited by Martin G. Möhrle and Ralf Isenmann. Berlin: Springer Berlin Heidelberg; Springer, 2005: 141–160.

Specht, Dieter; Mieke, Christian. "Weitsicht durch Analyse: Das Technologie-Roadmapping profitiert von der Patentanalyse als Informationsquelle." *wissenschaftsmanagement* 3, May/June (2004): 21–25.

Suh, Jong H; Park, Sang C. "Service-oriented Technology Roadmap (SoTRM) using patent map for R&D strategy of service industry." *Expert Systems with Applications* 36, 3, Part 2 (2009): 6754–6772.

Teece, David J; Pisano, Gary. “The Dynamic Capabilities of Firms: an Introduction.” *Industrial and Corporate Change* 3, no. 3 (1994): 537–556.

Teece, David J; Pisano, Gary; Shuen, Amy. “Dynamic capabilities and strategic management.” *Strategic Management Journal* 18, no. 7 (1997): 509–533.

Vecchiato, Riccardo; Roveda, Claudio. “Strategic foresight in corporate organizations: Handling the effect and response uncertainty of technology and social drivers of change.” *Strategic Foresight* 77, no. 9 (2010): 1527–1539.

Wasserman, Stanley; Faust, Katherine. *Social network analysis: Methods and applications*. 1st ed. Cambridge: Cambridge University Press, 2007.

Weiss, Sholom M. *Text Mining: Predictive methods for analyzing unstructured information*. New York: Springer, 2010.

Yoon, Byungun; Phaal, Robert; Probert, David R. “Morphology analysis for technology roadmapping: application of text mining.” *R&D Management* 38, no. 1 (2008a): 51–68.

Yoon, Byungun; Phaal, Robert; Probert, David R. “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, 2008b: 417–422.

3 The Role of Media in Innovation Systems

Abstract: While innovation is recognized as an outcome of systemic interaction, the elements enabling such interaction are often neglected in scientific debate. Based on the literature on innovation systems, this article attempts to study the role of media as a central actor in creating a public sphere where innovation discourse can take place. This study focuses on knowledge flows and systemic interactions and, therefore, examines in detail the link between science and the public. Publication data is a commonly used source for examining scientific developments. However, it is not only scientific achievement per se which is relevant. In addition, the issues passing through to other areas of the innovation system to reach the public are important. Therefore, news reporting is used as an indicator here. As part of the media system, news is a recognized channel for innovation diffusion and plays an important role in society. Thus, the aim of this article is to argue the benefit of integrating the media in the innovation system debate and to develop a methodology to automatically compare scientific and media discourses. Therefore, a text mining framework has been developed. The results deliver valuable input for examining the present state of themes and technologies and, thereby, support future planning activities.

Keywords: Innovation System, Text Mining, Foresight, Media Analysis, Publications Analysis, Future Technology Analysis

3.1 Introduction

Insights in innovation systems and their dynamics and architecture are relevant for future planning due to the close link between foresight, policy planning, and the performance of innovation systems (Alkemade et al., 2007). Therefore, an in-depth analysis of current developments is crucial for capturing the state-of-the-art as a starting point to build future assumptions and strategies. Besides trend recognition, this relates to the long-term observation of thematic and technical developments to gain insights for future orientation. Moreover, an analysis should not only address one area (e.g., science) but should also consider the larger framework (innovation system) in order to emphasize systemic interaction and to contribute to the evidence that innovation is systemic. As a matter of fact, the elements in innovation systems are not independent of each other—there is a continuous flow of knowledge between them. Thus, it is important to examine the intersections and links between different parts of these systems. However, the elements enabling these interactions have not been sufficiently addressed in the existing literature.

Based on the current literature on innovation systems, this article proceeds with the assumption of media being a central actor, enabling a public sphere where innovation discourses can take place. Thus, apart from its role in science, policy, and the economy, media should be considered in terms of its societal functions and role in the spread of innovation. After introducing an adapted innovation system model as conceptual framework, the methodology will focus on the link between science and the media. Publication data is a commonly used source for examining scientific progress (e.g., Leydesdorff and Milojević, 2015). However, developments in science are not the only relevant factor; it is also noteworthy which issues pass to other areas of the innovation system. News reporting can be used as an indicator in this regard. As part of the media system, news is an important channel for the spread of innovation and it has certain key functions in society. News contributes to the formation of public opinion and is an important channel for the diffusion of innovation (Rogers, 1995). However, news has rarely been used in this context until now. Currently, there is no methodology for the (automatic) comparison of news articles and scientific publications but theoretical discussions (e.g., Franzen et al., 2012).

With the assumption that media is an intermediary in innovation systems, the objective of this article is to develop a methodology to automatically compare scientific and media discourses to map systemic interactions or (technical) changes based on *textual data*. It is examined if differences in the discourse of science and media can be measured and mapped based on news articles and scientific publication abstracts. Therefore, a framework based on text mining is developed. This might deliver insights in the spread and diffusion of concepts and the chronological order of events. To test and illustrate the methodology, three topics driven by different angles are used—*cloud computing*, *artificial photosynthesis,* and *vegan diet*. The differences in these three cases may highlight the strengths and weaknesses of the methodology.

This article starts with a description of the basic building blocks, i.e., innovation system, foresight, and the (societal) role of the media, in Section 1.2. Then, Section.1.3 describes the framework of analysis while Section 1.4 introduces the three case studies. In Section 1.5, the results are discussed and final conclusions are drawn.

3.2 Foundations

This following points out the meaning of innovation, foresight, and innovation systems, with a focus on mass media and its impact on innovation and change.

3.2.1 Innovation System and Foresight

Innovation and change are an outcome of systemic interaction. This non-linear process includes many feedback loops and is considered in its national (Freeman, 1987), regional (Cooke, 2001), sectoral (Malerba, 2002), and technological contexts (Bergek et al., 2008). Definitions of innovation systems highlight how the interplay of institutions influences technology and innovation (Freeman, 1987) and innovation systems are described as *"[…] all important economic, social, political, organizational, institutional, and other factors that influence the development, diffusion, and use of innovations (Edquist, 1997)"*. These definitions emphasize the role of diffusion and interaction; therefore, the dynamics of these systems are most important. Among others, Hekkert et al. (2007) describe functions of innovation systems to measure system performance and dynamic interactions. These functions, such as *knowledge diffusion* or *market formation*, are important in assessing the performance of the system. On the other hand, understanding innovation systems and their dynamics and architecture is most relevant for future planning activities due to the close association between foresight, policy planning, and the performance of innovation systems (Alkemade et al., 2007). In this article, foresight is understood as a structured discourse about possible and plausible futures involving the relevant stakeholders. One of the basic assumptions underlying foresight is that the future is not predictable. However, thinking about possible future developments and related consequences may influence the present decisions that affect our future. Therefore, an in-depth analysis of current (technological) developments and their spread and societal acceptance is crucial. In principle, future technology analysis (FTA) and foresight can assist in reorienting and improving innovation systems and bringing together different stakeholders and actors (e.g., Martin and Johnston, 1999).

Aligning innovation system functions with FTA, the contribution of FTA (related to innovation policy) lays in "*[…] providing safe spaces for new ideas to emerge and existing knowledge* *to be combined in novel ways (Cagnin et al., 2012)".* This leads to a better understanding of future challenges and broadening of the knowledge base in decision making. Therefore, foresight may also serve as a framework for analysis. Apart from the debate on contributions of foresight to the analysis of the innovation system, the argument to consider foresight as a systemic process is strengthened. As Andersen and Andersen (2014) point out, foresight requires a systemic understanding because, otherwise, the impact of foresight is limited due to its weak conceptual understanding. So the context (innovation system) and the current dynamics should be taken into consideration for meaningful foresight.

3.2.2 Mass Media, its societal Role, and its Impact on Innovation

While it is commonly agreed that innovation needs to be viewed systemic, the society as a framework or mass media as a distribution channel are no explicit elements of prominent definitions of innovation systems (Waldherr, 2012, 2008). For this reason, this article discusses the role of the society and especially the media as diffusion channel and positions them in the innovation system debate.

The media contributes to our knowledge about the world (e.g., Luhmann, 2009). As a matter of fact, mass media has certain functions in society (Burkart, 2002). The most crucial one is the *information function*, which relates to neutral knowledge transfer as well as to influencing the formation of public opinion. The media distributes selected information to which it adds its own interpretation or version of truth (e.g., Kabalak et al., 2008). In addition, the media has a *critique and control function* in democratic societies, for scientific results as well (Franzen et al., 2012). Therefore, media mirrors public discourse and its evolution to a certain degree (see Stauffacher et al., 2015 for a comparable case).

As a matter of fact, media discourse may influence innovation processes (e.g., Waldherr, 2008). For instance, by reporting about new innovations and technologies, the media can influence and attract attention. Additionally, the media can influence public opinion by commenting on innovation (critique function of media). Furthermore, media has a recognized role in innovation diffusion (Rogers, 1995; Schenk, 2012; Karnowski, 2013). However, the literature on innovation systems does not acknowledge media's role as an intermediary between different actors, its functions in society, or its meaning for the spread of innovation. This article attempts to analyze the dynamics and processes of diffusion, in which media is recognized as a crucial element. Therefore, the next section introduces an adapted model.

3.2.3 The conceptually adapted Innovation System Model

Waldherr (2012) argues that mass media is an important intermediary in the triangle of politics, economy, and research (see Figure 3-1). Mass media enables public communication, while society is seen as the overall framework with three subsystems: economy, politics, and science. The link between media and the political system comprises political factors that influence the media. Further on, there is an exchange of money and attention between media and the economy, while media reputation is primarily relevant for firms. Additionally, economic actors learn about changing societal norms, values, and interests through media. And science needs public attention to build legitimacy and reputation.

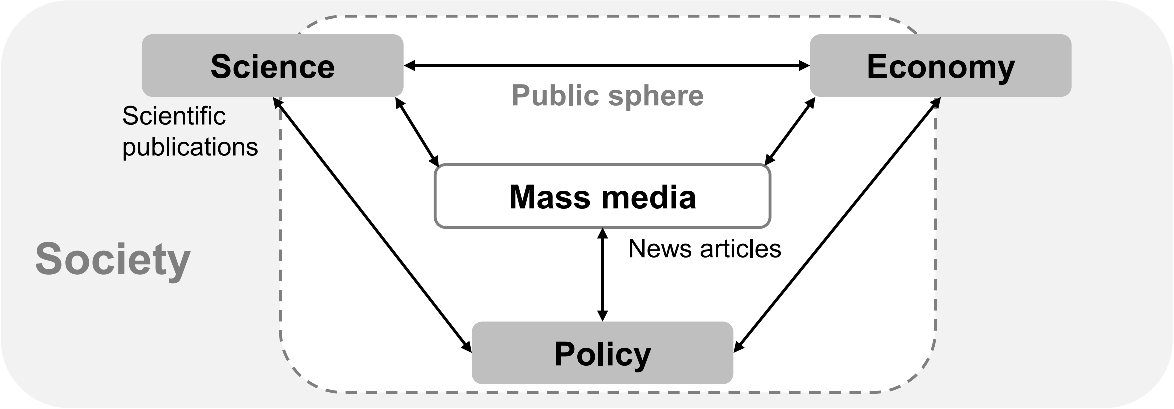


Figure 3-1 Adapted innovation system model (own illustration with reference to Waldherr, 2012)

Although this model is on a high aggregation level, it illustrates the core dependencies very well. Therefore, this model serves as a conceptual framework for the methodological part of this article on knowledge flows and dynamics in innovation systems. Later sections of this article focus on the relation between research and the media to examine when and in what form it is written about certain issues. This builds upon newspapers and scientific publications as sources of data. Moreover, apart from the structural description, innovation system functions may guide this analysis. At least three of the seven functions identified by Hekkert et al. (2007) demonstrate why mass media should be regarded as part of the innovation system model (Waldherr, 2012). These three functions are described in the following and related to the examination of the link between science and media.

Knowledge diffusion through networks: The exchange of information is seen as an essential function of networks (*learning by interaction* or *learning by using*). For example, policy decisions need to be in sync with the latest technological developments, which, in turn, should be in line with current norms and laws. This includes transfer of knowledge from science, politics, and economy to broad societal areas and, likewise, the coordination of research results with changing social norms and values (Waldherr, 2012; Hekkert et al., 2007). As stated above, media is an important diffusion channel between different elements of the innovation systems. To map this function, media discourse and the interrelation between societal and scientific discourse may be analyzed. For example, this function is mapped by the number of workshops, conferences, and research collaborations for a specific topic (Alkemade et al., 2007).

Guidance of the search and selection process: Since resources are scarce, this function describes the selection process, which includes focusing on specific applications. Therefore, mass media serves as a forum for the different stakeholders to negotiate the direction of technological change (Waldherr, 2012). This function underlines the fact that technological change needs to relate to external requirements and is not autonomous from the rest of the system. Additionally, this enables the alignment of technological developments with societal needs and the media can play an intermediary role in this process. The analysis of this function relates to mapping targets set by governments or the number of articles addressing the technology (Hekkert et al., 2007).

Creation of legitimacy: The spread of knowledge about an innovation is crucial to achieve societal legitimacy. Thus, this function addresses actions influencing the acceptance of an innovation in the manner of interest groups or lobbies. To map this function, the sentiment of news articles might be examined to see if certain views are pushed or provided with an intention.

3.3 Methodology: Comparing Datasets

The following section demonstrates how to compare the public and scientific discourse by assessing the knowledge flow between science and the public. The method builds on scientific publications and news reporting. This section begins with a description of the preliminaries of publication analysis and media analysis as methodological base for this article, after which the analysis framework is introduced. Three cases illustrating the methodology are outlined in Section 4.

3.3.1 Publication Analysis

Scientific publications describe the output of scientific work, thus providing a means to measure and assess scientific activity and performance. The statistical analysis of the publication data related to a specific theme or technology reveals insights on aspects such as trends, developments, and new research directions (see Leydesdorff and Milojević, 2015 for an overview). Publication analysis generally uses different data fields (e.g., year of publication, keywords, and abstracts) depending on the research interest. This article carries out text mining on the abstracts of the publications as summaries of the articles. This decision reduces the cleaning effort that is higher for full articles. Moreover, the text length of the abstract is comparable to the second type of data source—news articles.

Text mining is frequently used in publication analysis (Cunningham et al., 2006; Kostoff, 2012). This includes applications analyzing title, abstracts, and keywords (e.g., Glänzel, 2012) but also full texts (e.g., Glenisson et al., 2005). Concerning mapping of (technological) changes, most articles build on co-word analysis (Leydesdorff and Welbers, 2011; van Eck et al., 2010). For example, Cobo et al. (2011) examine the thematic evolution of a research field and concentrate on co-word analysis in combination with some performance indicators (e.g., number of publications, citations, and h-index). This article has a different focus; it not only describes the topic but also maps the differences in the datasets and the chronological order of terms to capture the dynamics in the topic (systemic interactions). With regard to the comparison of datasets, some previous work has compared patent and publication data (e.g., Daim et al., 2006). But, to the knowledge of the author, none has compared the content of publication abstracts and news articles. There are studies on the general relation between science and mass media (e.g., Franzen et al., 2012). However, this article especially has a methodological focus and attempts to develop a framework analyzing the interactions between science and media (with the example of publication abstracts and news articles).

3.3.2 Analyzing News Articles

Primarily, news articles are editorially checked (*controlled* content), cover a broad spectrum of themes, and have a clear language (e.g., few spelling mistakes and proper sentences). Thus, the text quality of news articles is comparable to scientific publications. In addition, news articles have a clear time stamp (date of publication). Moreover, news articles are archived and can be retrieved from databases such as *LexisNexis*. *LexisNexis* is still a popular source for analyzing media discourse across the world. In recent years, alternatives such as *Google News* have emerged (e.g., Weaver and Bimber, 2008 for a comparison) and newspaper archives are available online (e.g., *Der Spiegel*, *Die ZEIT*, and *The New York Times*). There are some known reliability and validity problems with *LexisNexis* and digital news archives in general (Deacon, 2007). However, using *Google News* involves a higher effort in searching, storing, and processing the articles. In addition, some forms of content such as images are not relevant for this article. Therefore, data from *LexisNexis* is used in the following.

Content analysis is the core method for analyzing news articles and is defined as “*[…] a research technique for making replicable and valid inferences from texts (or other meaningful matter) to the contexts of their use (Krippendorff, 2013)”.* Text is divided by its key features that are coded using a variable schema. Discourse analysis uses methods from content analysis and examines text elements that are part of a larger discourse. Thereby text elements are studied in terms of their relation to each other. However, manual screening and coding are not adequate for larger datasets, which is why automatic approaches have gained relevance (O'Connor and Banmann, 2011). Classic approaches cannot process the required volume of data, which often leads to reduction in the sample size due to resource problems (Scharkow, 2012). In fact, in recent years, more and more applications have emerged with regard to text mining. For example, Pollak et al. (2011) examine contrasting patterns in news articles from the UK, USA, and Kenya. By comparing local and Western media they study ideological aspects and press coverage. Holz and Teresniak (2010) identify changes in topics on the basis of the New York Times corpus by computing the co-occurrence of terms over time. Of course, these automatic methods are also criticized; first, that automated content analysis will never be able to replace careful reading (Grimmer and Stewart, 2013) and, second, for the potential loss in meaning (Sculley and Pasanek, 2008). However, text mining delivers summaries and reduces the costs and effort involved in analyzing large text collections and will therefore be used in the following.

Sentiment analysis is often used for the automatic detection of opinions and attitudes in texts. It is a classification problem where each text is treated as a unit that needs to be classified based on the words it contains (positive, negative, or neutral). In recent years, the research effort spent on sentiment analysis has increased. For example, an overview is given in Ravi and Ravi (2015). Normally, sentiment analysis is applied on subjective texts such as movie reviews or web forums (Li and Wu, 2010). In this article, sentiment analysis has been considered for being applied on news articles. Newspapers express opinions that can usually be analyzed. However, a literature review revealed that sentiment analysis is difficult to apply on news articles. When sentiment analysis is applied on news, the scope must be clearly defined (Balahur and Steinberger, 2009). Also the views or perspectives on the article, such as intention of author or reader interpretation, needs to be distinguished. The *source* of the opinion is emphasized to be the journalist or the newspaper (in most cases), but the *target* is more difficult to distinguish (e.g., distinguishing good and bad news from good and bad sentiment). So even for reported facts, judging good or bad news depends on one's perspective and differs individually. As a further point, news articles cover larger subject domains compared to e.g., product reviews. This makes it even harder to (automatically) identify the *target* (Balahur et al., 2013). Additionally, opinions are expressed less explicitly and more indirectly in the news than in other texts. Owing to these reasons, this article does not attempt to apply automatic sentiment analysis.

3.3.3 Introducing the technical Framework

The technical framework developed in this article is based on Python and SQL. As described above, two data sources have been used, namely *Web of Science* for the scientific publications and *LexisNexis for the news articles*. For the export of news articles, a filter has been set on English newspapers (e.g., *The New York Times*, *The Guardian*). Effort has been spent on converting the database output of *LexisNexis* to a machine-readable format. To this end, a customized Python module has been programmed, which automatically identifies the key fields (e.g., heading and publication date) and extracts the main text of the article. Additionally, duplicate articles and articles containing fewer than 50 words have been deleted.

First, the number of records per year in the two datasets, news articles and scientific publications, has been compared. This shows if there has been any media attention to the topic at all, how extensive the debate is, and if, in principle, it can be assumed that people have learnt something about the topic. Second, the texts are analyzed in more detail as described in the following. This is motivated by the question of which aspects the datasets focus on, the coverage and volume of reporting, who reported first, and whether there are recognizable influence directions.

Text pre-processing and noun phrase extraction

This step structures the texts (abstracts and news articles) and transforms them to a numeric dataset. Nouns are separately extracted from the texts to summarize and structure the content for machine processing. First, each text is broken into single words. *Part-of-speech* tags are assigned to the words of each sentence to describe their grammatical instance (Bird et al., 2009). To extract noun phrases from each sentence, regular expressions are formulated by filtering out single nouns or chains of nouns (e.g., *carbon dioxide*, *interoperability*). Lemmatization on plural forms and a thesaurus (to match varying spellings such as American and British English and replace abbreviations) are used for cleaning. Additionally, a *stopword* list removes very common phrases such as *paper* or *study*. The single texts are short, so using binary frequencies of the terms in each document is sufficient. Finally, the resulting numeric data is stored in the SQL-database for further processing.

Matching and comparing datasets

Several SQL tasks are conducted to compare and match the datasets. Term networks for the 100 most frequent terms are drawn for each dataset as an initial overview of the terms and their co-occurrence. The networks illustrate how the terms are interconnected and, therefore, dependent on each other. In contrast to wordclouds, terms occurring together in a document are linked. Additionally, the graph metrics and graph sorting algorithms (here: *force atlas*) give additional input. The node size depends on the binary frequency of a term in the dataset and not, as in other applications, on the node degree. Frequency is a suitable measure due to the fact that the density and connectivity are normally high in term networks and, otherwise, all nodes are of equal size. Comparing two networks gives an orientation with regard to the ongoing discussions and summarizes the content. The networks are visualized with Gephi (Bastian et al., 2009).

Next, publication abstracts and news articles are matched to identify common and unique terms. Technically, this is realized in SQL by comparing term frequencies and occurrence.

The results are mapped as *pie bubble charts*. Terms are depicted as bubbles. These bubbles contain pie charts that have sections for each dataset. The size of the sections shows the relative frequency of the term in each dataset. For example, the term *security* is present in 55% of the news articles and in 11% of the abstracts. The bubble size relates to the summed relative frequency of a term per dataset. For each term, the size of the term is the sum of the binary term document frequency (*tdf*) per dataset, calculated by:

$$size(term) = tdf_{abstract}(term) + tdf_{news}(term)$$

This means that large bubbles represent more frequent terms than smaller bubbles. While the bubbles are randomly distributed on the y-axis (avoiding overlaps of bubbles), the x-axis represents the degree of inclusion in the news (left side) or in scientific publications (right side). It is calculated by:

$$x = \frac{tdf_{abstract}(term) - tdf_{news}(term)}{tdf_{abstract}(term) + tdf_{news}(term)}$$

So the difference between the *tdf* of the abstract minus the *tdf* in the news is divided by the size of the bubble (the summed relative frequency per dataset). The *pie bubble charts* enables a comparison of the substantive orientation of the datasets. It may also indicate special terminology, especially when terms only occur in one dataset, such as *ingredient* in the case of *vegan diet*.

The common terms are analyzed for their first occurrence in the dataset (chronological order). This step is based on a SQL-query. This shows time differences and may indicate drivers for development and changes.

3.4 Case Studies and Results

This section describes three cases—*cloud computing*, *artificial photosynthesis,* and *vegan diet*. These very different cases were deliberately chosen to illustrate the methodology and highlight differences. It is commonly acknowledged that *cloud computing* has huge application potential and market relevance. In contrast, *artificial photosynthesis* is a (basic) research topic and relatively few public discourses are expected on this topic. The third case, *vegan diet*, is a temporary societal phenomenon of changing nutrition habits.

For all three cases, data has been retrieved from *Web of Science* and *LexisNexis* by a keyword-based search. The first search with regard to *cloud computing* has been restricted to articles because, otherwise, the output is very large (more than 10.000 results); for the other two cases, articles and proceedings have both been searched. Table 3-1 describes the data and gives an overview on the searches and their results.

		Scientific publications *Web of Science*	**News articles** *LexisNexis*
Cloud computing	**Search string**	*TS = ("cloud computing")* *Articles*	*"cloud computing"* *Newspaper articles*
	Time	2007 - 2014	2007 - 2014
	Size of dataset	2630 entries; 2578 abstracts	420 news articles
Artificial photosynthesis	**Search string**	*TS = ("artificial photosynthesis")* *Articles + Proceedings*	*"artificial photosynthesis"* *Newspaper articles*
	Time	1990 - 2014	1980 - 2014
	Size of dataset	1407 entries; 1326 abstracts	407 news articles
Vegan nutrition	**Search string**	*TS = (vegan)* *Articles + Proceedings*	*"vegan"* *Newspaper articles*
	Time	1990 - 2014	1990 - 2014
	Size of dataset	507 articles; 492 abstracts	721 news articles

Table 3-1 The three datasets

3.4.1 Cloud Computing

Cloud computing (e.g., Armbrust et al., 2010) is an emerging technology linked to core managerial implications, which leads to new modes of IT service offering. In short, it can be described as decentralized storage and computing services. Its strong management aspect emphasizes that data distinct from scientific publications is relevant to measure the spread and change of this topic. For *cloud computing*, data from 2007 to 2014 has been retrieved. In all, 2630 articles were retrieved, of which 2578 had an abstract. In addition, 420 news articles were downloaded. Figure 3-2 gives an overview on peaking or declining attention. In the first three years, media and science have addressed the issue equally and the numbers develop in parallel. From 2010 onwards, the media attention has decreased continuously, while scientific publication numbers have increased up to over 900 records in 2014.

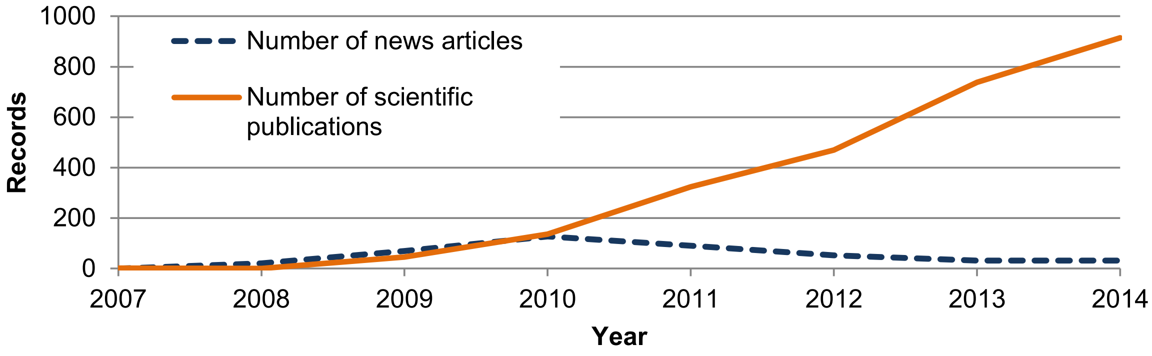
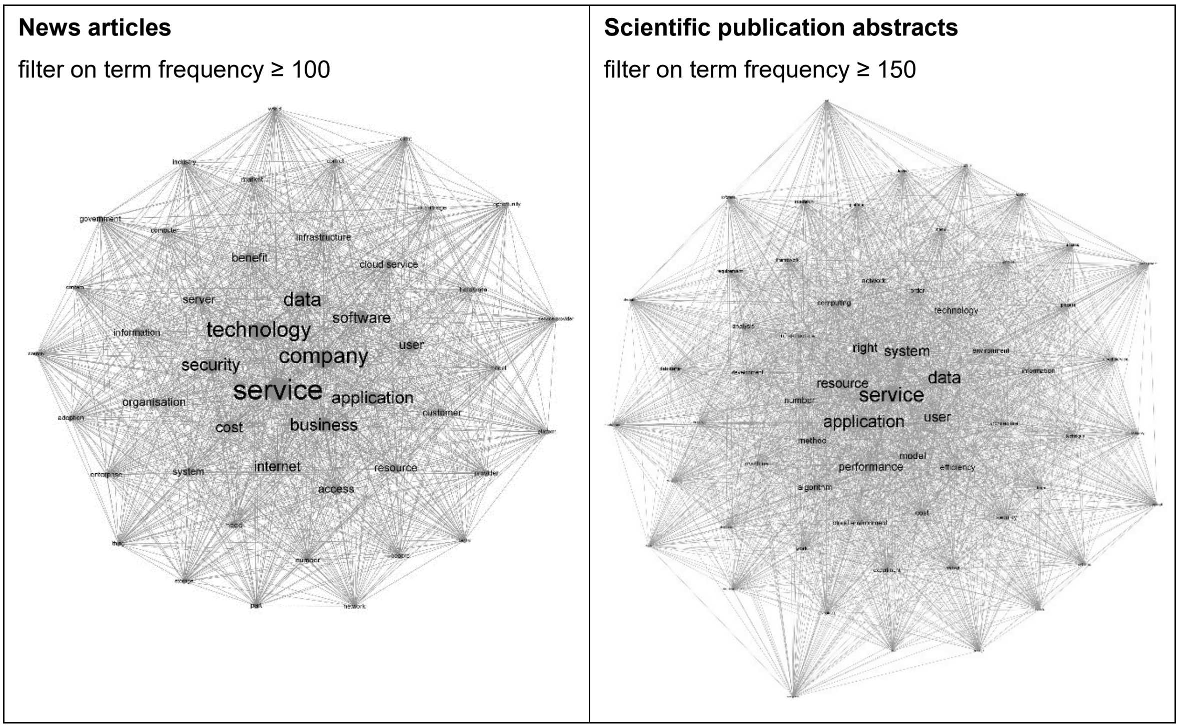


Figure 3-2 Cloud computing: number of records

In the next step, the texts are processed and the content is summarized in term networks for an overview. These term networks illustrate the links among the 100 most frequent terms. As Figure 3-3 shows, both term networks highlight *service, data,* and *application,* but they are linked differently. In the news, they are frequently mentioned together with *company*, *security,* or *business;* this underlines the management and business focus. In contrast, in the abstracts these terms are closely linked to *system*, *performance, efficiency,* and *resource*. This indicates that the scientific discourses are more about computing while the news reports more on the market aspects (e.g., *organization*, *cost*).


News articles

filter on term frequency ≥ 100

Scientific publication abstracts

filter on term frequency ≥ 150

Figure 3-3 Cloud computing: network of terms

Next, the *pie bubble chart* directly compares the frequency of terms in the two datasets (Figure 3-4). For example, *algorithm*, *method,* and *experiment* are much more frequent in scientific abstracts. *Data* and *application* frequently occur in both datasets. On the other hand, *company*, *business,* and *storage,* as well as *market*, *enterprise,* and *customer,* are more frequent in the news. This underlines the fact that news articles are more management-driven for describing organizational structures (e.g., *director*), while the abstracts contain typical scientific vocabulary (e.g., *fault tolerant*, *scheme*, and *simulation result*). Obviously, the news reports have a business and market focus (e.g., *cost*, and *benefit)*. *Security* is more frequent in the news than in the abstracts; possibly because *security* affects the acceptance of *cloud computing* in enterprise environments.

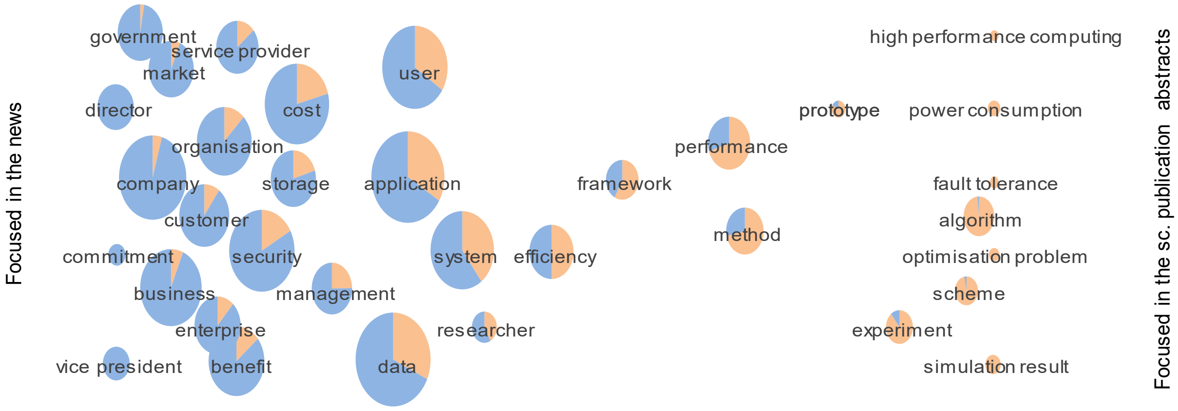


Figure 3-4 Cloud computing: pie bubble chart (selection of terms)

Almost no term occurs in abstracts first and in the news later. For example, *service-oriented architecture* was first mentioned in the news in 2008 and in abstracts in 2011. Obviously, the news reports before scientific publications get published.

When interpreting these results, several points should be kept in mind concerning the comparison of the two datasets. First, a scientific review process needs more time than publication of news articles. This leads to a time delay in the first occurrence of terms in the abstracts and is evident in the case of cloud computing. Second, research results anticipating outcomes are often additionally published in the news (e.g., researchers giving interviews; reports about ongoing research). Third, the news generalizes (e.g., *technology and data*) and tends to use fewer specific or technical terms (e.g., *virtual machine* and *map reduce*). Finally, the news might pick up a specific term or trend from other newspapers and reports a lot about it. In contrast, scientific publications specifically address research gaps, potentially leading to less repetition of terms. The last two points explain why the terms occurring frequently in the news are larger in Figure 3-4 than the terms focused in the publication abstracts (e.g., *high performance computing* and *fault tolerant).* This observation recurs in the second case, *artificial photosynthesis*.

3.4.2 Artificial Photosynthesis

Artificial photosynthesis deals with energy generation from sunlight and holds potential as a regenerative source of energy (see, e.g., House et al., 2015 for an overview). Research in this field is still at a basic level despite going on for more than 40 years. Back in the 1980s, there were already initial news articles reporting on the potential of this technology. The following analysis focuses on the time period from 1990 to 2014. In all, 1407 scientific articles were retrieved from *Web of Science,* (1.326 of these featured an abstract) and 407 news articles from *LexisNexis*. As Figure 3-5 depicts, there have been relatively few news articles until 2005, while the number of scientific publication is slightly higher. This indicates a limited public discourse, even as the number of scientific publication grew steadily, especially from 2010 onwards. The scientific activity rose from 79 records in 2010 to 272 records in 2014, while there is still a lag in media attention (around 49 reports per year on average from 2010 to 2014).

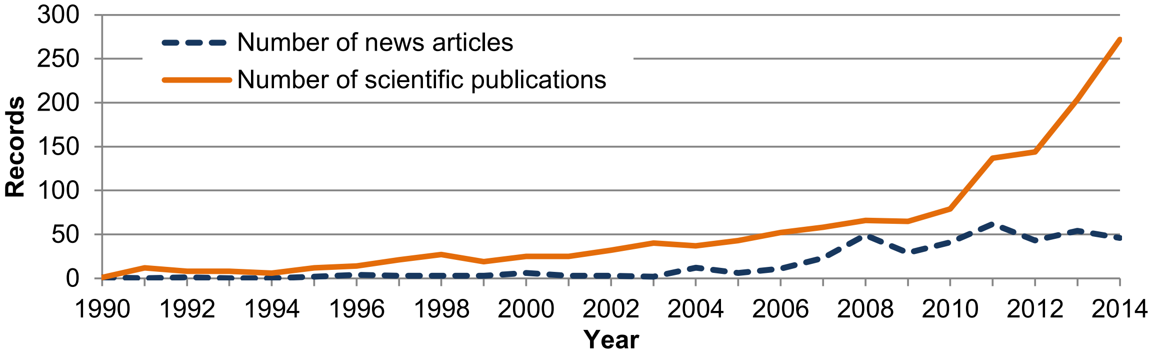
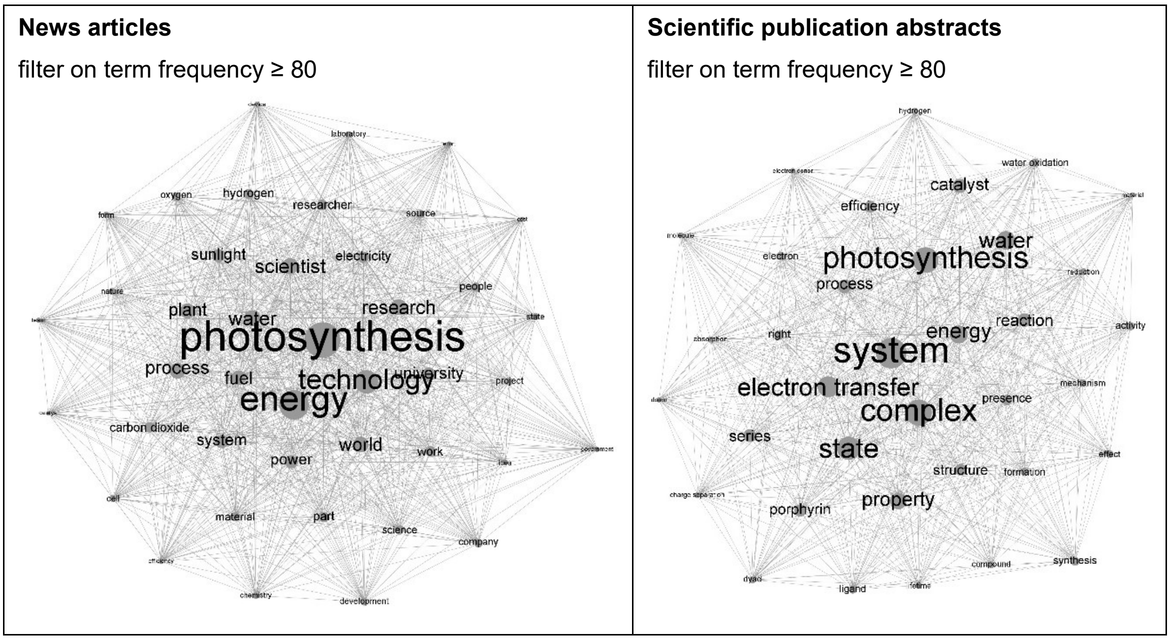


Figure 3-5 Artificial photosynthesis: number of records

Next, the texts are processed. The 100 most frequent terms per dataset are visualized in term networks (Figure 3-6). As the comparison of the two networks shows, the focus of the news lays on *photosynthesis* for energy generation. It seems as if they report a lot about scientific work (e.g., *research*, *university*, and *scientist*). In contrast, the scientific abstracts use more scientific vocabulary (e.g., *complex*, *electron transfer*, *catalyst, and reaction*).


News articles

filter on term frequency ≥ 80

Scientific publication abstracts

filter on term frequency ≥ 80

Figure 3-6 Artificial photosynthesis: network of terms

Then, the data is illustrated as a *pie bubble chart*. As Figure 3-7 shows, terms occurring only in the abstracts are rare (e.g., *phenyl, fluorescence spectra*), with the exception of *electron transfer* and *water oxidation*. Frequent terms such as *professor* and *university* occur only in the news. This indicates that the news reports a lot about scientific work and progress. The abstracts are dominated by (scientific) terms such as *absorption and oxidation*. Thematically, news concentrates on reporting about research results and energy generation.

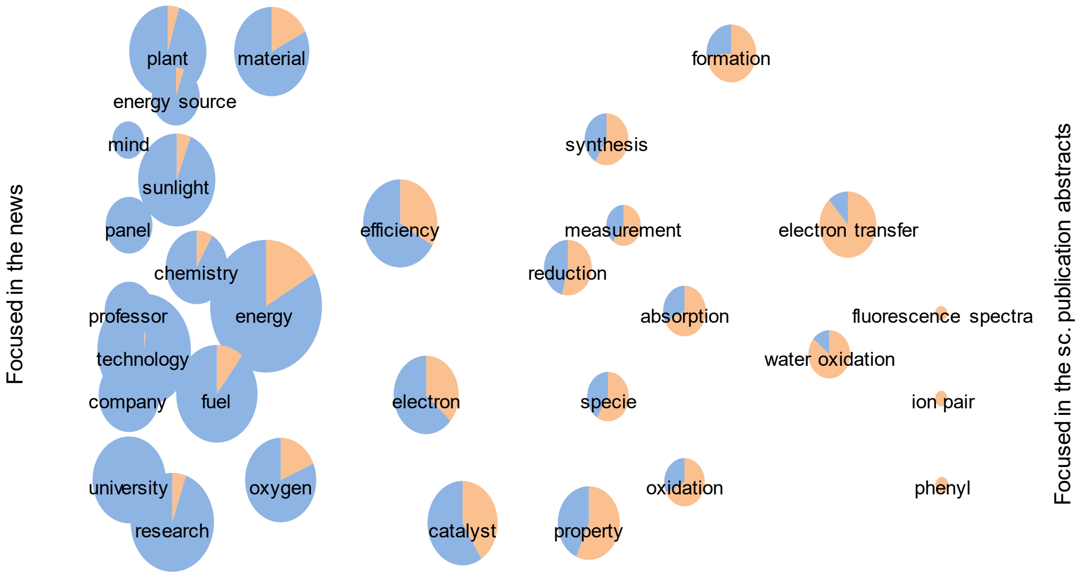


Figure 3-7 Artificial photosynthesis: pie bubble chart (selection of terms)

Concerning the chronological order and first occurrence of the terms in the two datasets, there is a set of scientific terms that first occurs in the abstracts, such as *conversion* (1991 compared to 2001 for news), *absorption* (abstract: 1991; news: 2008), or *synthesis* (abstract: 1991; news: 1995). But other terms such as *semiconductor* (abstract: 1996; news: 1992) first occur in the news. One observation from the previous two steps is that the news mostly reports about scientific work and discovery, but there seems to be a certain delay for some topics.

3.4.3 Vegan Diet

Vegan diet has become a (societal) trend in recent years. This type of diet that is free of meat and animal products has been attracting more and more followers. Compared to the other two cases, this topic is assumed to be more society-driven and less influenced by scientific discoveries. It is not an actual technology but rather a change in behavior that might showcase a *social change* and thus be more visible in news reporting. Data has been retrieved from 1990 to 2014 (Figure 3-8). In all, 507 articles have been downloaded from *Web of Science*, of which 492 include an abstract. On the other hand, 721 news articles were retrieved from *LexisNexis*. From 2004 onwards, more has been published on this topic in the news than scientific publications. This may be related to the societal hype of the vegan diet and the public attention it attracts.

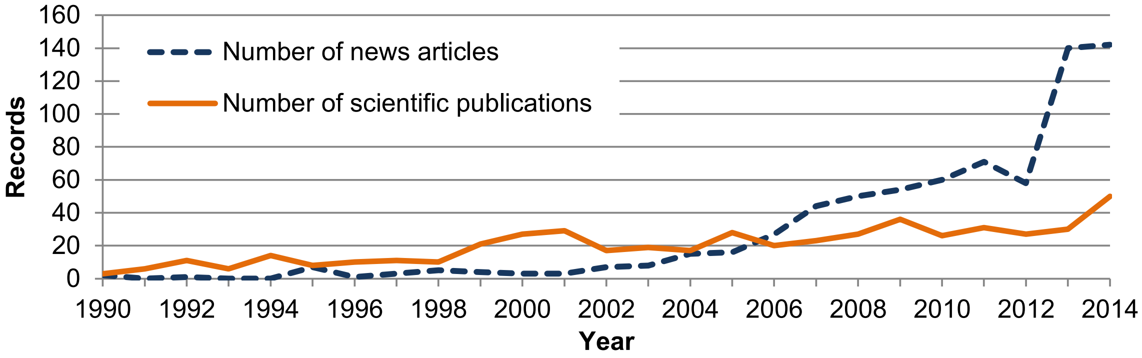


Figure 3-8 Vegan diet: number of records

The texts (news articles and abstracts) are processed and for an overview on the thematic focus in each dataset, term networks are drawn (Figure 3-9). For the networks, *vegan*, *diet*, and *vegan diet* are excluded from this step because they are very frequent and part of the search strategy. Obviously, the news concentrate on *food*, *people*, and the names of different diets (e.g., *veganism* and *vegetarian*). Additionally, *milk* and *dairy* as well as *meat* and *animal* are frequently mentioned. Therefore, the focus is on lifestyle and diet. *Health*-related issues play a subordinate role, as opposed to the scientific discourses which report a lot on the health impact of the vegan diet and signs of possible deficiency (e.g., *intake*, *effect*, and *differences*). Thus, most abstracts describe medical experiments and statistics (e.g., *participant, sample*).

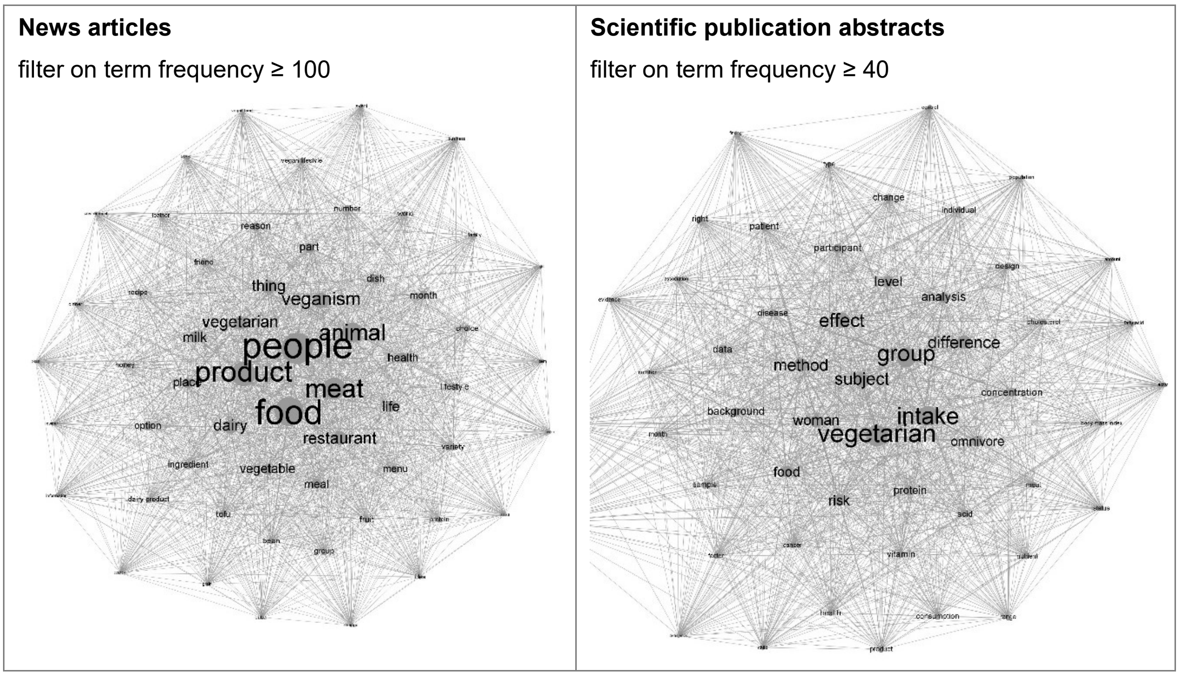


Figure 3-9 Vegan diet: network of terms

As Figure 3-10 illustrates, common terms are *food*, *diet,* and *vegetarian*. Additionally, the news reports a lot about types of grains (e.g., *grain* and *seed*). Obviously there is a difference between medical vocabulary used in the abstracts (e.g., *intake* and *fatty acid*) and food and nutrition issues in the news (e.g., *body weight*). This analysis indicates that science and the common public are talking about different things. Again, the results illustrate that the news focuses on lifestyle and cooking, while the abstracts mostly cover medical and health issues.

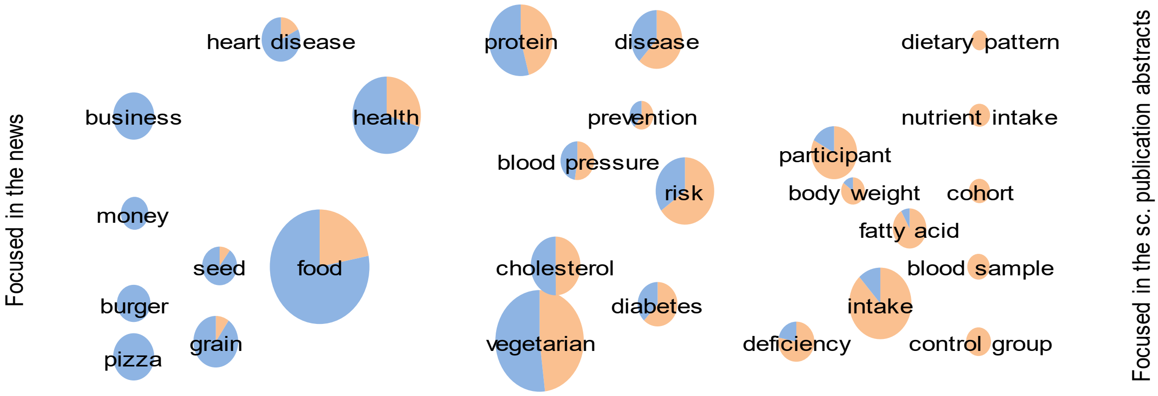


Figure 3-10 Vegan diet: pie bubble chart (selection of terms)

The comparison of the first occurrence of terms in the news and the abstracts again highlight that they cover different aspects, resulting in major time lags. This relates for example to *cereal* (news: 1992; abstract: 2000) and *grain* (news: 1990; abstract: 1997), or *risk factor* (news: 2004; abstract: 1992) and *protein intake* (news: 2007; abstract: 1991).

3.5 Discussion and Conclusion

The aim of this article was to argue the benefit of integrating the media in the innovation system debate and to develop a methodology to automatically compare scientific and media discourses using text mining. This section assesses the methodology and discusses its role in FTA and innovation systems.

As argued before, the media should be integrated in the innovation system debate because diffusion is emphasized in many definitions of innovation systems, plus the media has functions in society and plays a recognized role in innovation processes. On the example of the link between science and the media, this article tries to develop a method for the automatic comparison of scientific and media discourse where few work exists so far. While publication data is frequently used as an indicator for science and technology performance, quantitative examination of news articles is rarely applied in this context. As the three cases in Section 4 illustrate, the method developed here automatically summarizes textual content and visualizes it in different ways (term networks and pie bubble charts). This illustrates how the terms are connected and gives a quick overview on thematic focus in the two datasets. Thereby, the results describe thematic differences in scientific and media discourses (e.g., reporting about scientific results or lifestyle issues). Terms common in both datasets can easily be distinguished. Additionally, the diffusion of certain issues can be estimated, thus providing a solid starting point for future explorations.

In any case, a broader context is necessary for the interpretation and validation of the results, but they can trigger interesting discussions. Basically, this method is applicable to generate hypotheses on the evolution of a topic that should be tested and validated by additional methods. These forms of data analysis have certain inherent limitations and, therefore, should be combined with qualitative expert assessments (see e.g., Cozzens et al., 2010). In fact, some research questions require a more in-depth analysis. For example, sentiment analysis still needs to be done manually, and storylines in articles or political directions can hardly be examined automatically. However, more data can be processed with an automatic approach, even if it is only for a first orientation or for advance coding schemes for content analysis. Of course, the analysis grid in this article is coarse, but it gains relevance in times of increasing data volumes implying an increased reading effort. Today’s challenge is not in finding the right information but in extracting the relevant information to generate knowledge from the quantity (Montoyo et al., 2012). Therefore, certain mechanisms are needed. This method is an attempt to this end, especially in the context of foresight where the current state of technology needs to be captured at the beginning of almost every process.

This article examines if it is possible to automatically compare news articles and publication abstracts and develops a method for this purpose. After this first attempt proves that the research path followed in this article is promising, it can be expanded in future applications. This especially relates to four points. First, more complex text mining methods might be used. For this work, effort had been spent on processing and structuring the news articles. Clustering or classification (e.g., Pollak et al., 2011) are deliberately not used here because domain knowledge about a topic is necessary or the approach requires a high learning effort. However, this might be tried in future work. As a second point, different or more (textual) data can be used to address or emphasize different aspects of the innovation system. This relates, for example, to not only policy briefs, press releases, market figures, research funding calls, or newsletters, but also social media content. Third, according to Moore’s innovation lifecycle (Moore, 2006), the market penetration of an innovation is imminent after the media attention decreases. This theory is evident for *cloud computing* where 2010 is a turning point. An in-depth examination of this correlation was not a subject of this article but may be an interesting point for future research. So *technology lifecycles* might be examined on the basis of combinations of different data sources (e.g., social media, online news, and patents) with reference to known models. Fourth, additional (qualitative) methods might be used to validate the hypothesis and observations.

As stated before, foresight is context dependent; so the larger context (such as innovation systems) should be taken into account. Therefore, mapping the present is essential for the success of the whole foresight process (Andersen and Andersen, 2014) and the method developed here is valuable for the analysis of the current state of technology and ongoing dynamics. Additionally, it may recognize current trends to estimate future development paths. This delivers valuable insights for future technology analysis and foresight. Further on, with regard to foresight and innovation, foresight still lacks a clear theoretical base (Fuller and Loogma, 2009; Öner, 2010) though it might have stronger links to innovation studies. Both innovation and foresight can be considered at different levels (*micro* to *meso*) and more effort should be spent on (theoretically) linking them in future work.

As shown previously, it is reasonable to integrate the media as an element in innovation systems due to the fact that media has functions in society and its role in innovation diffusion. As a consequence, the innovation system model has been adapted in this article to emphasize interaction and diffusion. However, the model introduced in Section 2.3 is highly aggregated. For an in-depth analysis, the innovation system needs to be described more precisely. For instance, this means to take structural, national or technological differences into account and formulate the three areas (policy, economy, science) in more detail. However, the aim of this article is to develop a methodology to capture dynamics at the intersection of science, media and society rather than examining structural differences. Section 2.3 introduces three system functions and another aspect is to examine if these functions can be aligned with the method developed here.

First, the exchange of information and knowledge transfer related to *knowledge diffusion* can, to a certain degree, be mapped. In addition, differences between scientific and media debate can be illustrated. Principally, the intensity of the media reporting varies and also what they are writing about. For *cloud computing*, the media distributed much *knowledge*, but though its interest decreased after the first years (after 2010). On the other hand, in the case of *vegan diet,* the media reports a lot, but about different things. More exchange is noticeable for the first two cases than for the last one (vegan diet). However, as Hekkert and Negro (2009) conclude, many of these knowledge diffusion processes are not explicitly noticeable and therefore cannot be mapped and recognized.

Second, g*uidance of the search* is difficult to map explicitly as well. Of course, the number of articles can be mapped, but if they raise specific expectations is difficult to say by this kind of analysis. Everything around *selection process* and *priority setting* is difficult to extract automatically. And, what also holds for the third point *creation of legitimacy*, as the literature review in Section 3.2 showed, sentiment and opinions are too complex to extract automatically from news articles. It is difficult to assess (automatically) which interest group is reporting, who is influencing the report, or if positive or negative opinion on a technology is expressed. Summarized, this methodology can principally support analysis of the dynamics, but a direct assignment to the functions is strained. Generally, in terms of development, diffusion, and adoption of technologies as primary goal of innovation systems, the results of this method allow certain conclusions, but as indicated before, there remains a great deal on the level of hypothesis that should be proven by additional examinations. In addition, the generated databases of news articles and scientific publications (as another result of the method) can be used for additional (qualitative) analyses such as *event process extraction* as applied by Negro (2007) or Tigabu et al. (2015).

This article lays a basis that can be developed in various directions. The results are promising and the method should be developed further, for example, by using different data sources or applying different data analysis. The results deliver an overview on differences in orientation (e.g., management, scientific reporting, lifestyle issues) and intensity of reporting, leading to hypotheses and starting points for further (more detailed) explorations. In fact, text can be used to measure and model dynamics in innovation systems and more effort should be spent here in future. Finally, automatic approaches for a quick overview of large datasets are relevant in our present time of increasing volume of data.

References

Alkemade, Floortje; Kleinschmidt, Chris; Hekkert, Marko P. “Analysing emerging innovation systems: a functions approach to foresight.” International Journal of Foresight and Innovation Policy 3, no. 2 (2007): 139–168.

Andersen, Allan D; Andersen, Per D. “Innovation system foresight.” *Technological Forecasting and Social Change* 88 (2014): 276–286.

Armbrust, Michael; Fox, Armando; Griffith, Rean; Joseph, Anthony D., et al. “A view of cloud computing.” *Communications of the ACM* 53, no. 4 (2010): 50–58.

Balahur, Alexandra; Steinberger, Ralf. “Rethinking Sentiment Analysis in the News: from Theory to Practice and back.” Proceedings of the '1st Workshop on Opinion Mining and Sentiment Analysis'. Seville, Spain, 2009.

Balahur, Alexandra; Steinberger, Ralf; Kabadjov, Mijail; Zavarella, Vanni, et al. “Sentiment analysis in the news.” *arXiv preprint arXiv:1309.6202 (*2013): 2216–2220.

Bastian, Mathieu; Heymann, Sebastien; Jacomy, Mathieu. “Gephi: An Open Source Software for Exploring and Manipulating Networks.” Conference: Proceedings of the Third International Conference on Weblogs and Social Media. San Jose, California, USA, 2009.

Bergek, Anna; Jacobsson, Staffan; Carlsson, Bo; Lindmark, Sven; Rickne, Annika. “Analyzing the functional dynamics of technological innovation systems: A scheme of analysis.” *Research policy* 37, no. 3 (2008): 407–429.

Bird, Steven; Klein, Ewan; Loper, Edward. *Natural language processing with Python*. 1st ed. Beijing, Cambridge [Mass.]: O'Reilly, 2009.

Burkart, Roland. *Kommunikationswissenschaft: Grundlagen und Problemfelder*. 4th ed. Wien [et al.]: Böhlau, 2002.

Cagnin, Cristiano; Amanatidou, Effie; Keenan, Michael. “Orienting European innovation systems towards grand challenges and the roles that FTA can play.” *Science and Public Policy* 39, no. 2 (2012): 140–152.

Cobo, M. J; López-Herrera, A. G; Herrera-Viedma, E; Herrera, F. “An approach for detecting, quantifying, and visualizing the evolution of a research field: A practical application to the Fuzzy Sets Theory field.” *Journal of Informetrics* 5, no. 1 (2011): 146–166.

Cooke, Philip. “Regional Innovation Systems, Clusters, and the Knowledge Economy.” *Industrial and Corporate Change* 10, no. 4 (2001): 945–974.

Cozzens, Susan; Gatchair, Sonia; Kang, Jongseok; Kim, Kyung-Sup, et al. “Emerging technologies: quantitative identification and measurement.” *Technology Analysis & Strategic Management* 22, no. 3 (2010): 361–376.

Cunningham, Scott W; Porter, Alan L; Newman, Nils C. “Special issue on tech mining: Tech Mining: Exploiting Science and Technology Information Resources.” *Technological Forecasting and Social Change* 73, no. 8 (2006): 915–922.

Daim, Tugrul U; Rueda, Guillermo; Martin, Hilary; Gerdsri, Pisek. “Forecasting emerging technologies: Use of bibliometrics and patent analysis.” *Technological Forecasting and Social Change* 73, no. 8 (2006): 981–1012.

Deacon, David. “Yesterday's Papers and Today's Technology: Digital Newspaper Archives and 'Push Button' Content Analysis.” *European Journal of Communication* 22, no. 1 (2007): 5–25.

van Eck, Nees J; Waltman, Ludo; Noyons, Ed C. M; Buter, Reindert K. “Automatic term identification for bibliometric mapping.” *Scientometrics* 82, no. 3 (2010): 581–596.

Edquist, Charles, ed. *Systems of innovation: Technologies, institutions, and organizations*. London: Pinter, 1997.

Franzen, Martina; Rödder, Simone; Weingart, Peter. “Wissenschaft und Massenmedien: Von Popularisierung zu Medialisierung.” In *Handbuch Wissenschaftssoziologie,* edited by Sabine Maasen, Mario Kaiser, Martin Reinhart and Barbara Sutter. Wiesbaden: Springer Fachmedien Wiesbaden, 2012: 355–364.

Freeman, Christopher. *Technology, policy, and economic performance: Lessons from Japan*. London, New York: Pinter Publishers, 1987.

Fuller, Ted; Loogma, Krista. “Constructing futures: A social constructionist perspective on foresight methodology.” *Futures* 41, no. 2 (2009): 71–79.

Glänzel, Wolfgang. “Bibliometric methods for detecting and analysing emerging research topics.” *El Profesional de la Informacion* 21, no. 2 (2012): 194–201.

Glenisson, Patrick; Glänzel, Wolfgang; Janssens, Frizo; de Moor, Bart. “Combining full text and bibliometric information in mapping scientific disciplines.” *Information Processing & Management* 41, no. 6 (2005): 1548–1572.

Grimmer, Justin; Stewart, Brandon M. “Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts.” *Political Analysis (*2013): 1–31.

Hekkert, Marko P; Negro, Simona O. “Functions of innovation systems as a framework to understand sustainable technological change: Empirical evidence for earlier claims.” *Technological Forecasting and Social Change* 76, no. 4 (2009): 584–594.

Hekkert, Marko P; Suurs, Roald A.A; Negro, Simona O; Kuhlmann, Stefan; Smits, Ruud E.H.M. “Functions of innovation systems: A new approach for analysing technological change.” *Technological Forecasting and Social Change* 74, no. 4 (2007): 413–432.

Holz, Florian; Teresniak, Sven. “Towards Automatic Detection and Tracking of Topic Change.” edited by Alexander Gelbukh and Alexander Gelbukh. Proceedings of the 11th International Confrence on Computational Linguistics and Intelligent Text Processing: Springer, 2010: 327–339.

House, Ralph L; Iha, Neyde Yukie Murakami; Coppo, Rodolfo L; Alibabaei, Leila, et al. “Artificial Photosynthesis: Where are we now? Where can we go?: Where are we now? Where can we go?” *Journal of Photochemistry and Photobiology C: Photochemistry Reviews (*2015).

Kabalak, Adam; Priddat, Birger P; Rhomberg, Markus. “Medien als Schnittstelle zwischen politischen und ökonomischen Strukturen - Politische Kommunikation in der Perspektive der Institutionenökonomie.” In *Massenmedien als politische Akteure,* edited by Barbara Pfetsch and Silke Adam. Wiesbaden: VS Verlag für Sozialwissenschaften, 2008: 52–70.

Karnowski, Veronika. “Diffusionstheorie.” In *Handbuch Medienwirkungsforschung,* edited by Wolfgang Schweiger and Andreas Fahr. Wiesbaden: Springer Fachmedien Wiesbaden, 2013: 513–528.

Kostoff, Ronald N. “Text mining for science and technology - a review part I – characterization/scientometrics.” *Scientometrics* 1, no. 1 (2012): 11–21.

Krippendorff, Klaus. *Content analysis: An introduction to its methodology*. 3rd ed. Los Angeles, London: SAGE, 2013.

Leydesdorff, Loet; Milojević, Staša. “Scientometrics.” In *International encyclopedia of the social & behavioral sciences*. 2nd ed., edited by James D. Wright. Amsterdam: Elsevier, 2015: 322–327.

Leydesdorff, Loet; Welbers, Kasper. “The semantic mapping of words and co-words in contexts.” *Journal of Informetrics* 5, no. 3 (2011): 469–475.

Li, Nan; Wu, Desheng D. “Using text mining and sentiment analysis for online forums hotspot detection and forecast.” *Decision Support Systems* 48, no. 2 (2010): 354–368.

Luhmann, Niklas. *Die Realität der Massenmedien*. 4th ed. Wiesbaden: VS, Verlag für Sozialwissenschaften, 2009.

Malerba, Franco. “Sectoral systems of innovation and production: Innovation Systems.” *Research policy* 31, no. 2 (2002): 247–264.

Martin, Ben R; Johnston, Ron. “Technology Foresight for Wiring Up the National Innovation System.” *Technological Forecasting and Social Change* 60, no. 1 (1999): 37–54.

Montoyo, Andrés; Martínez-Barco, Patricio; Balahur, Alexandra. “Subjectivity and sentiment analysis: An overview of the current state of the area and envisaged developments.” *Decision Support Systems* 53, no. 4 (2012): 675–679.

Moore, Geoffrey A. *Crossing the chasm: Marketing and selling disruptive products to mainstream customers*. New York, NY: Collins Business Essentials, 2006.

Negro, Simona O. “Dynamics of technological innovation systems: the case of biomass energy.” Netherlands Geographical Studies 356, 2007.

O'Connor, Brendan; Banmann, David. “Computational Text Analysis for Social Science: Model Assumptions and Complexity.” *public health* 41, no. 42 (2011): 1–7.

Öner, M. A. “On theory building in Foresight and Futures Studies: A discussion note.” *Futures* 42, no. 9 (2010): 1019–1030.

Pollak, Svenja; Coesemans, Roel; Daelemans, Walter; Lavra, Nada. “Detecting contrasting patterns in newspaper articles by combining discourse analysis and text mining.” *Pragmatics (*2011): 647–683.

Ravi, Kumar; Ravi, Vadlamani. “A survey on opinion mining and sentiment analysis: Tasks, approaches and applications.” *Knowledge-Based Systems* 89 (2015): 14–46.

Rogers, Everett M. *Diffusion of innovations*. 5th ed. New York: Free Press, 1995.

Scharkow, Michael. *Automatische Inhaltsanalyse und maschinelles Lernen*. Berlin: epubli GmbH, 2012.

Schenk, Michael. *Medienwirkungsforschung*. 3rd ed. Tübingen: Mohr Siebeck, 2012.

Sculley, D; Pasanek, Bradley. M. “Meaning and mining: the impact of implicit assumptions in data mining for the humanities.” *Literary and Linguistic Computing* 23, no. 4 (2008): 409–424.

Stauffacher, Michael; Muggli, Nora; Scolobig, Anna; Moser, Corinne. “Framing deep geothermal energy in mass media: the case of Switzerland.” *Technological Forecasting and Social Change* 98 (2015): 60–70.

Tigabu, Aschalew D; Berkhout, Frans; van Beukering, Pieter. “The diffusion of a renewable energy technology and innovation system functioning: Comparing bio-digestion in Kenya and Rwanda.” *Technological Forecasting and Social Change* 90 (2015): 331–345.

Waldherr, Annie. “Gatekeeper, Diskursproduzenten und Agenda-Setter — Akteursrollen von Massenmedien in Innovationsprozessen.” In *Massenmedien als politische Akteure,* edited by Barbara Pfetsch and Silke Adam. Wiesbaden: VS Verlag für Sozialwissenschaften, 2008: 171–195.

Waldherr, Annie. “The Mass Media as Actors in Innovation Systems.” In *Innovation Policy and Governance in High-Tech Industries,* edited by Johannes Bauer, Achim Lang and Volker Schneider. Berlin, Heidelberg: Springer, 2012: 77–100.

Weaver, David A; Bimber, Bruce. “Finding News Stories: A Comparison of Searches Using Lexisnexis and Google News.” *Journalism & Mass Communication Quarterly* 85, no. 3 (2008): 515–530.

4 Using Twitter for Foresight: An Opportunity?

Abstract: Since its foundation, Twitter has established as a popular micro-blogging service and platform for public real-time communication. Principally, users *tweet* their ideas and opinions and share information in up to 140 characters. This broad spectrum of content attracts the research interest of different scientific disciplines with diverging research focus and interests such as human communication behavior or trend predictions. Concerning foresight, the value of Twitter has not been discussed or examined yet. While the use of Web 2.0 and social media at the intersection to foresight is addressed in a range of articles, none exclusively focuses on Twitter. Here this article concentrates on and considers different applications and use cases to reveal how to use Twitter in foresight exercises. After a short introduction to Twitter and its basic principles, an analysis framework is introduced and illustrated for the case of *#quantifiedself*. Additionally, options are outlined how to interact with a global network of people using Twitter as communication platform. As the results show, both, the monitoring of topics and technologies, but also the active user engagement, are supported. In summary, Twitter is a beneficial path to follow in foresight and future planning processes as an opportunity to extend the considered data sources and to increase the number of involved stakeholder views.

Keywords: Twitter, Foresight, Text Mining, Social Media, Web Mining

4.1 Introduction

Twitter has established as a worldwide micro-blogging service and platform for public real-time communication (Bruns and Burgess, 2012; Java et al., 2007). As both a social network and an information-sharing platform, Twitter offers real-time news and covers a broad spectrum of topics. Twitter thereby aggregates many opportunities for conducting scholarly research, and so has attracted rising attention in recent years. Obviously, each scientific discipline has diverging interests. Whereas some study human communication behavior and social networks (e.g., Marwick and Boyd, 2011), others perform trend predictions (e.g., Asur and Huberman, 2010) or observe the online communication during crisis and natural disasters (e.g., Terpstra et al., 2012). A lot of promising work from other disciplines has been published, but the potential of Twitter for foresight is rarely discussed. In general, the use of Web 2.0 and social media is addressed in a range of articles on future policy planning or trend recognition (e.g., Cachia et al., 2007; Haegeman et al., 2012; Grubmüller et al., 2013). However, the use of Twitter as an information source or platform for foresight exercises is rarely considered, although a lot of options actually exist.

These observations lead to the research question addressed in this article: Does Twitter have any potential to be used in foresight? This relates, on the one hand, to Twitter as a data source and what can be (automatically) retrieved from it. Therefore, a framework is developed to illustrate the benefit. On the other hand, Twitter as a social media platform enables active engagement and user interaction; some examples are described.

This article begins with introducing Twitter and its basic principles in Section 4.2. After describing the scientific discourse about foresight and Web 2.0 in general, this article examines the opportunities that arise from Twitter in Section 4.3. At this point, first applications are outlined on Twitter as information source or platform in foresight. Finally, the results are discussed and conclusions are drawn in Section 4.4.

4.2 Twitter: An Overview

The following introduces Twitter and its key characteristics, the principles of Twitter analysis and an overview on Twitter as a research field.

4.2.1 Key Characteristics

Twitter was launched in October 2006 and has become the largest micro-blogging service since then with currently 500 million tweets per day (Twitter, 2015). According to a recent statistic, around 22% of the worldwide internet users are active on Twitter (Globalwebindex, 2014). This article concentrates on Twitter because of its broad spectrum of covered aspects, the contained web links to additional content, the global spread and because it provides real-time access to user-generated content. Compared to other services Twitter is not only designed for disseminating news but also for active user engagement and an exchange of messages as tweets.

In particular, Twitter has five functionalities: tweets, hashtags, @-messages, retweets, and follower relations (see Table 4-1 for an overview). Each user can publish tweets and subscribe to the tweets of other users by f*ollowing* them. This creates a social network of users and follower relationships as a “directed friendship model (Marwick and Boyd, 2011). This is in contrast to the undirected models of other social media platforms as, for example, Facebook. Each tweet can be forwarded as a retweet, be directed to other users by *@-messages*, or annotated by a *#hashtag*. Additionally, the tweets can contain web links referring to, for example, news articles, press releases, or reports.

Tweet	As a message of 140 characters, tweets can contain @-messages, links and #hashtags. The tweets can be answered and retweeted by other users.
@-message	To mention other Twitter users in a tweet, their username is tagged with @.
#hashtag	By the #-symbol, terms are tagged and connected with tweets using that same term.
Retweet	By retweeting a message, it is forwarded to the user’s followers and can be shared within its network.
Follower	A follower *follows* other users on Twitter and thereby subscribe to the tweets of other users. Follower networks can be built up.

Table 4-1 Overview on basic functionalities of Twitter

Certainly, Twitter contains a lot of banal chatter. The main types of interaction are daily chatter and conversations, news reporting and information sharing (Java et al., 2007). However, Bruns and Burgess (2012) emphasize the role of social media channels such as Twitter in today’s public communication as being used first primary in private communication, but this has changed within the last years. Social media and Twitter are meanwhile increasingly used by politicians and organizations for communicating with their consumers or citizens. Moreover, Twitter developed from sharing mostly personal information to sharing diverse information (see, e.g. Risse et al., 2014).

To access this debate is most interesting for foresight, in particular to engage with different groups and stakeholders. So, in recent years, Twitter has established and a wide range of applications evolved as, for example, in enterprise-related communication (e.g., Stieglitz and Krüger, 2014), during crises and disasters (e.g., Terpstra et al., 2012), or in scholarly communication (e.g., Holmberg and Thelwall, 2014). Therefore, Twitter should be taken seriously and its potential, especially for foresight, will be examined in the following.

4.2.2 Twitter Data Analysis

No common way has prevailed how to conduct Twitter data analysis while there are qualitative and quantitative endeavors (see, e.g., Bruns and Burgess, 2012). In the course of this article, two strategies for getting data are distinguished as denoted in Figure 4-1: searching data (1) or active engagement (2). The former depends on a search strategy (as described below) while the latter depends on launching a discussion (see Section 4.3.3 for examples). This is followed by data gathering and data analysis while an interpretation of the results is generally the last step.

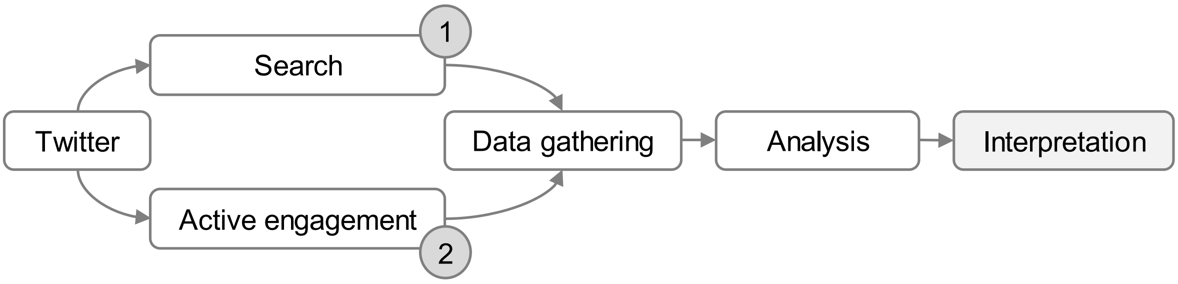


Figure 4-1 Twitter analysis: static search or active engagement

Twitter can be searched for terms, hashtags or (groups of) users. The complexity of this search depends on the individual field characteristics and how well the field can be delimited. For example, to search emerging technologies in general is more complex than searching for specific technologies such as *#quantifiedself* or *#bioeconomy* where certain hashtags are repeatedly used.

Several commercial and non-commercial tools exists for gathering and analyzing Twitter data (see for an overview e.g., Gaffney and Puschmann, 2014). Which tool to apply depends on the individual process requirements. Furthermore, it is possible to implement an own analysis framework using the Twitter API. Primarily, the Twitter API was designed for integrating Twitter in other web services and applications but is now also used for data gathering. In principal, Twitter asserts the monopoly right on its data because its business model relies on selling this data; naturally resulting in a conflict of interests (see, e.g., Puschmann and Burgess, 2014). So by most tools data can only be retrieved from now on.

The analysis of Twitter data reveals aspects as communication patterns, recent trends, user statistics, or follower networks. Apart from the tweet, metadata as tweet ID, geo coordinates of sender and the user ID are included in the data retrieved by the Twitter API. A qualitative tweet analysis delivers first insights, but with an automated approach more data can be processed. When analyzing Twitter data the handling and interpretation of retweets needs to be clarified. According to Metaxas et al. (2014), retweets express interest, trust or agreement. Boyd et al. (2010) describe retweets as form of validation and engagement with other users. This implies a certain relevance, but tools that automatically retweet on certain terms or hashtags reduce the expressivity of the received retweets. So for the course of this paper, tweets and retweets were distinguished and retweets were interpreted as *received attention*.

4.2.3 Twitter Research

There is an ongoing scientific discourse on Twitter and its analysis in different scientific disciplines as indicated by the following bibliometric analysis**.** Thereby, it is examined which scientific disciplines are involved, what they specifically address and how they are interconnected. For getting a rapid overview on their research interests, author keywords are analyzed.

Publication data (both articles and proceedings) related to Twitter[1] was extracted from the *Web of Science*-database (2.581 results). Figure 4-2 depicts an increasing publication activity within the last 9 years. For a comparison, data on the related social media platforms *Facebook* and *YouTube* and on *social media* was retrieved. *Social media* as generalization has the highest number of publications while *Facebook* has more and *YouTube* less publications than *Twitter*.

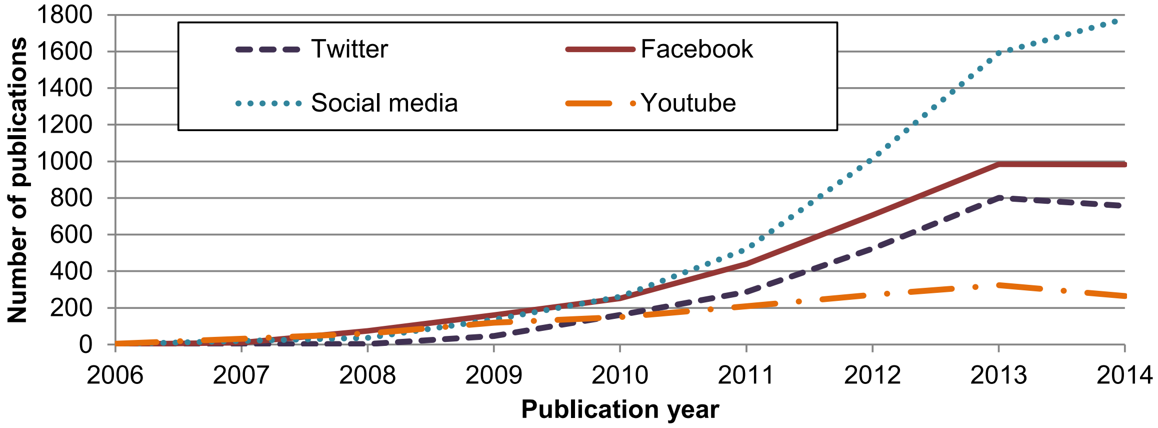


Figure 4-2 Twitter and related platforms (source: *Web of Science*; time interval: 2006 - 2014)

The data on the search for *Twitter* was retrieved for a more detailed analysis. Figure 4-3 contains the network of research areas active in the field of Twitter research. This network results from the mentioning of different disciplines related to a publication. Links denote connections between two disciplines and the size of the node depends on the number of linkages (node degree) and thereby the connectivity. The most active discipline is *computer science* involved in 47% of the considered articles and a focus on developing algorithms and improving data analysis methods. *Computer science* is strongly linked to *engineering* with an equal focus. While the bottom half of the network has a technical and engineering focus (e.g., chemistry, environmental sciences & ecology), other disciplines contribute to the field of Twitter research as well such as *business and economics* (e.g., trend prediction or brand communication) or the *social sciences* (focus on social networks and communication behavior). Furthermore, *psychology*, *neurosciences*, *pharmacology*, or *educational science* are active in the field of Twitter research,

This analysis supports the assumption of an (interdisciplinary) scientific exchange about Twitter. Remarkably, foresight and future planning do not occur in this analysis. As a consequence, possible applications will be a focus of this article and outlined in the next section.

[1] Search string: TS= “Twitter” AND DOCUMENT TYPES: (Article OR Proceedings Paper); Timespan: 2006-2014.

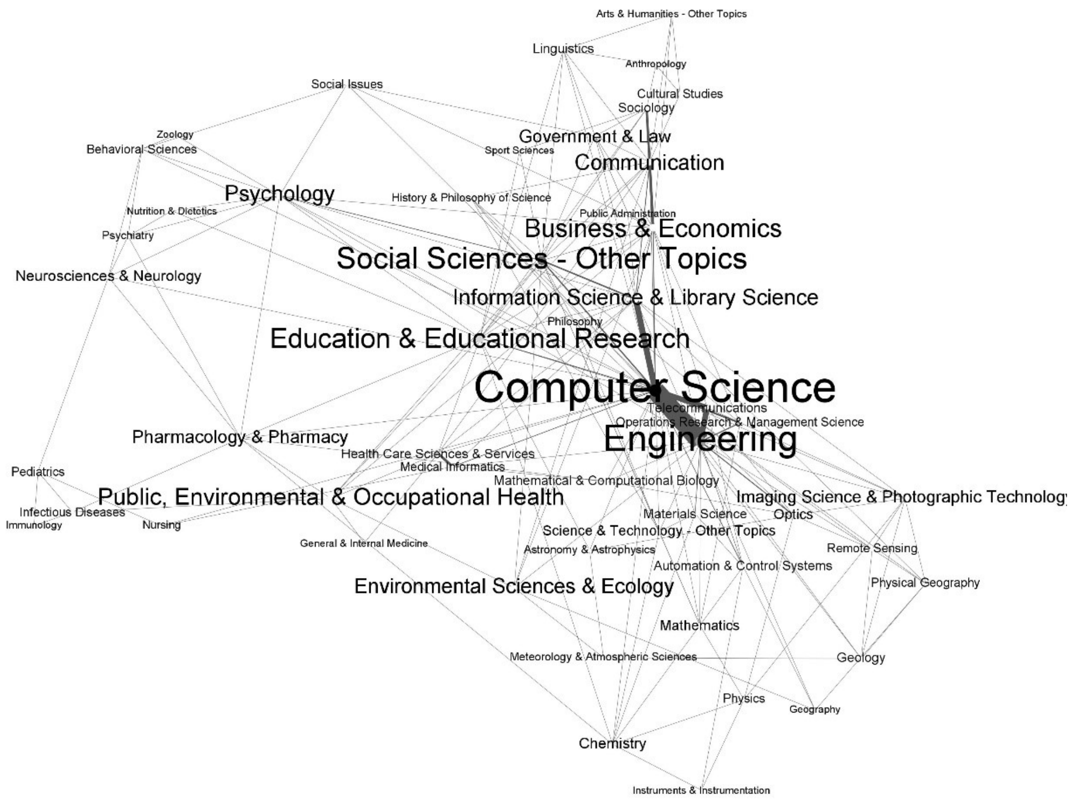

Arts & Humanities - Other Topics
Linguistics
Anthropology
Social Issues
Cultural Studies
Sociology
Zoology
Behavioral Sciences
Government & Law
Sport Sciences
Communication
Psychology
History & Philosophy of Science
Nutrition & Dietetics
Public Administration
Business & Economics
Psychiatry
Social Sciences - Other Topics
Neurosciences & Neurology
Information Science & Library Science
Philosophy
Education & Educational Research
Computer Science
Telecommunications
Operations Research & Management Science
Pharmacology & Pharmacy
Engineering
Health Care Sciences & Services
Medical Informatics
Pediatrics
Mathematical & Computational Biology
Public, Environmental & Occupational Health
Imaging Science & Photographic Technolog
Infectious Diseases
Materials Science
Optics
Immunology
Nursing
Science & Technology - Other Topics
General & Internal Medicine
Astronomy & Astrophysics
Remote Sensing
Automation & Control Systems
Environmental Sciences & Ecology
Physical Geography
Mathematics
Meteorology & Atmospheric Sciences
Geology
Physics
Geography
Chemistry
Instruments & Instrumentation

Figure 4-3 Network of research areas (node degree ≥ 4)

For a fast overview on the research interests, the 100 most frequent author keywords were analyzed and visualized as a term network (see Figure 4-4) using the method as described in Section 2. Central and well-connected nodes are *Twitter*, *social network* and *social media* – obviously covered in many abstracts. A cluster on analyzing Twitter data is located in the bottom of the network (e.g., *sentiment analysis*, *classification*, and *algorithm*). Furthermore, *politics*, *e-learning,* and *election* are contained in the network. This underlines that a breadth of research interests can be addressed using Twitter as a basis.

4.3 Opportunities with Twitter for Foresight

According to the understanding underlying this article, foresight is a structured dialogue about possible future developments among relevant stakeholders. This integrates qualitative and quantitative approaches where to build on with Twitter very well. The context where foresight is conducted (e.g., policy planning, corporate strategy) is not further distinguished in the course of this article because general options are discussed that do not require an exact situation.

The following section describes opportunities of Twitter for foresight – as a data source and as a platform for conducting foresight exercises. After summarizing the state-of-the-art, use cases are outlined and first experiments are conducted.

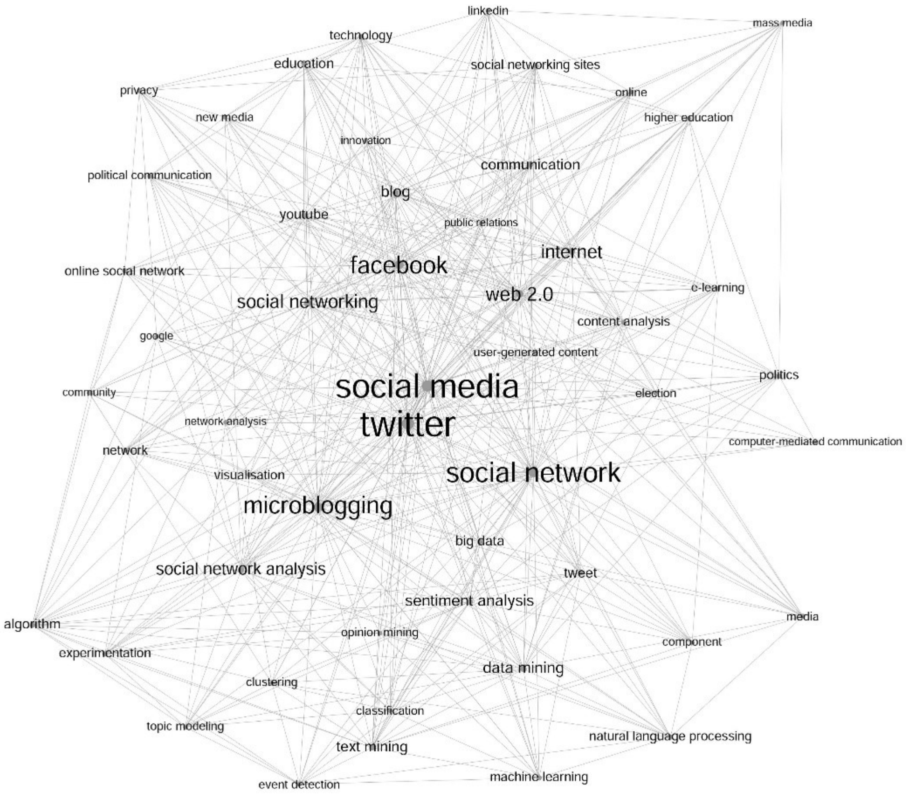


Figure 4-4 Network of author keywords (node degree ≥ 15)

4.3.1 Related Work on Foresight and Social Media

Related to the usage of Twitter in the context of foresight, little work was done so far. More research exists on the general application of Web 2.0 and social media in foresight and future-oriented planning. For example, Cachia et al. (2007) describe the potential role of online communities for foresight. They conclude that these communities hint towards changes and trends in sentiment and social behavior. Additionally, they encourage creativity and collective intelligence, serve for brainstorming and enable the collaborative development and debate of different future developments. As Gheorghiou et al. (2009) state, Web 2.0 enable new types of foresight exercise that are more interactive, make better use of online resources and develop content. In their work, they propose a Web 2.0 platform as extension of the Delphi method.

While companies use Twitter and Twitter monitoring for long (e.g., Stieglitz and Krüger, 2014), policy planning has just started. For example, Haegeman et al. (2012) emphasize the still limited use of Web 2.0 tools in policy planning. They tested a Web 2.0 framework for R&I priority setting. Grubmüller et al. (2013) examine social media (analysis) for future-oriented policy planning. They state that governments increasingly recognize the benefit of social media as information source and additionally as an instrument for receiving feedback and detecting future trends. In their project, they gathered citizen opinions to adapt future policy development. Apart from challenges as legal and ethical issues that need to be resolved, they conclude that social media are an ideal instrument to support policy planning.

Pang (2010) describes how to harvest online content produced by futurists for a *social scanning* framework where Twitter may be one part. Amanatidou et al. (2012) use Twitter in a horizon scanning framework for collecting web links on relevant issues. They further state that Twitter may be a source for identifying signals and examining their spread, but additionally the uptake by different communities. In addition, they see an interactive component in the use of Twitter. As their work already underlines, a set of different information sources needs to be combined in order to receive reliable results as each source has its bias. So among others they worked with Twitter, conducted a survey, processed foresight reports, and attended conferences. Schatzmann et al. (2013) outline some applications of Twitter summarized under the term *foresight 2.0*. However, most of these applications are intended for core trend predictions focused on the one future. This does not meet the aim of this article and its underlying foresight understanding.

Recently, altmetrics evolved at the intersection of scientometrics and social media trying to measure scientific activity apart from scientific publication behavior. Rather, altmetrics focuses on Twitter or other social media formats (Priem et al., 2012; Thelwall et al., 2013). For example, the scholarly communication of different disciplines on Twitter is examined (Holmberg and Thelwall, 2014). This of course requires that the community is online and contributes (tweets) together with the (reliable) identification of researchers on Twitter.

4.3.2 Twitter as a Data Source for Foresight

140 characters as for a tweet are small space – however, they can contain valuable information. Twitter cannot only be used for information-sharing but additional for searching information (e.g.,Teevan et al., 2011) and may deliver valuable input to examine societal change and concerns. Principally, most foresight exercises start with desk research to get an overview on current developments and previous activities regarding the topic. Here, Twitter as a platform can facilitate the search process, help to check *if* there is a public debate or any attention at all, *how* intense or emotional this debate is, which actors contribute and *what* is discussed.

In the following, a case study underlines the use of Twitter as a data source for foresight activities. Data was retrieved using the Twitter API and a framework for processing Twitter data was implemented based on Python and SQL. Tweets on a certain topic are read in together with the sender (as username), the number of retweets the tweet received, the date and the number of @mentions per user. This data is processed and the tweet is split into single words, @mentions, hashtags and web links. After cleaning the terms (e.g., matching word variants) and hashtag (e.g., merging plural and singular form), the data is stored in the database for being analyzed further.

The following describes the functionalities of this framework using the example of *quantified self*. In short, *quantified self* (e.g., Swan, 2012) is to incorporate technology, especially sensors and IT, into everyday life for self-monitoring of different parameters such as vital signs or nutrition. Table 4-2 summarizes descriptive parameters of the dataset restricted on English tweets (4.762 out of 6.284 tweets in total).

Search for	#quantifiedself
Time	2015-03-02 to 2015-04-14
Number of retweets	3.040
Number of tweets	1.722
Number of different users that tweet or retweet	1.657

Table 4-2 Key parameters of the dataset

Often, their remains the question of who the main actors are engaged with a topic. Principally, Twitter enables the observation of concurring actors, the classic competitors. In a more science oriented context, the research landscape can be observed by following other research institutes or conference news feeds. The number of followers and received @messages can indicate influential actors. Twitter-accounts with high follower numbers might be important organizations or users. For example, Table 4-3 shows the five most active users in the dataset on *#quantifiedself*. This table lists the number of followers an user has, the number of its tweets and retweets and the number of received @mentions. Thereby, this table indicates whom to follow if monitoring a topic. Additionally, their profiles can be checked for further information, regardless if they are organizations or private actors. As a further point, this analysis supports identifying relevant actors and experts that are sooner or later needed in most foresight exercises (e.g., workshops, interviews). Frequently mentioned users with a high number of tweets have a certain influence or meaning and might be contacted for the following foresight exercise for interviews, workshops, or surveys. Due to the real-time character of Twitter and its high actuality, this might reveal different actors than searching information elsewhere. So, one option is to analyze structural aspects of Twitter and identify relevant organizations or people tweeting on a certain topic.

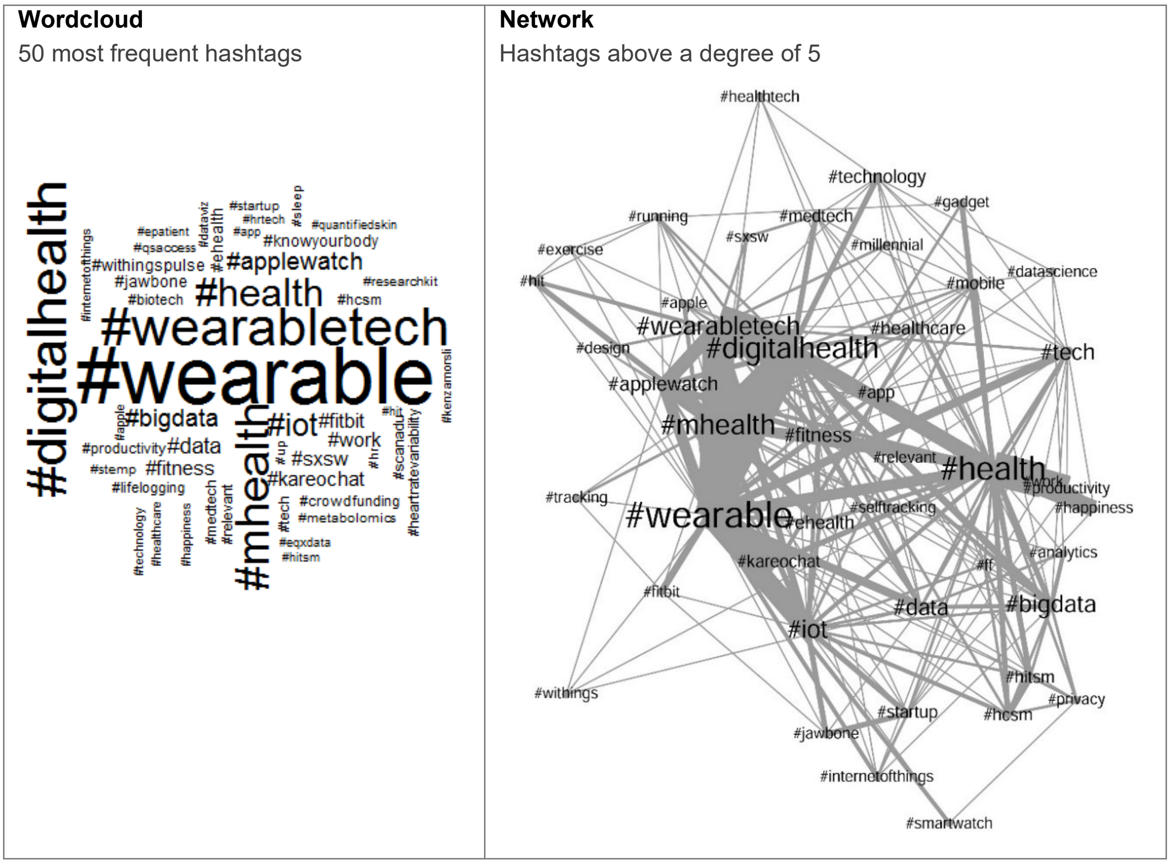
Username	**Followers**	**Retweet**	**Tweet**	**@mentions (no RT)**
eramirez	2960	53	78	34
jayfader	2173	2	27	3
mchiaviello	8313	0	23	1
quantifiedself	12709	4	24	20
agaricus	3884	3	16	3

Table 4-3 Top 5 users for *#quantifiedself*

Concerning the tweets, the most retweeted messages can be considered as a first orientation indicating aspects that get much attention. Further on, web links, hashtags and the messages' content give an overview about the ongoing debates and the content shared. To begin with, hashtags contained in the tweets were analyzed. Retweets were excluded for this analysis. The tweets are visualized as wordclouds (size of word depends on its frequency) and as networks (underlines which hashtags are frequently mentioned together and therefore connected).

As Figure 4-5 indicates, much in the field of *quantified self* is about wearables or health and its digital monitoring. As the wordcloud shows, products as the *apple watch* or *fitbit* are covered as well. In addition, it is examined how the hashtags are interconnected. As shown by the hashtag-network, *#health* is highly connected to *#wearable*, #digitalhealth or #bigdata. Additionally, *#digitalhealth* and *#wearabletech* or *#mhealth* and *#wearabletech* are strongly linked.

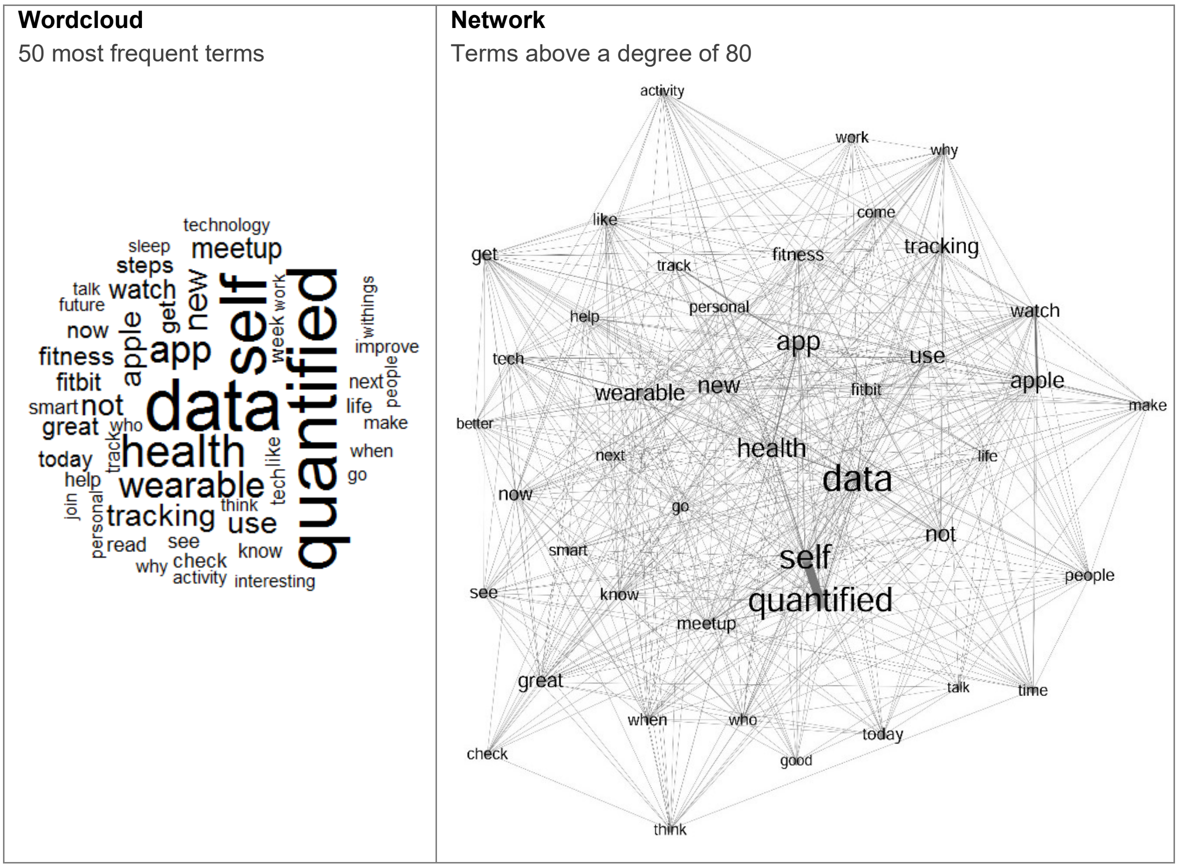
Next, the terms contained in the tweets were analyzed for further insights (see Figure 4-6). Here, stopwords were removed (e.g., *and*, *the*, *we*). For example, the term *data* is very frequent and, as the network shows, *data* is a highly connected term, closely linked to *app* and *health*.



Wordcloud
50 most frequent hashtags

Network
Hashtags above a degree of 5

Figure 4-5 Hashtags contained in the tweets on *#quantifiedself*



Wordcloud
50 most frequent terms

Network
Terms above a degree of 80

Figure 4-6 Analysis of tweets on *#quantifiedself*

The tweets on *#quantifiedself* contain 1.118 web links that can be checked for further input by applying web scraping (e.g., Russell, 2011). In this case, the website headers are retrieved and can be manually screened for relevant content (see Table 4-4). This content can contribute further insights, contribute timely issues and broaden the information base of the foresight exercise.

Contained link	***Website header***
http://t.co/6XuUGkqrIo	Is peer-reviewed science too slow to track wearable accuracy? | mobihealthnews
http://t.co/7Bv6SDkhC9	Defining a new indicator of cardiovascular endurance and fitness – Marco Altini
http://t.co/7KuN01teRR	Yasmin Lucero on Baby Tracking Quantified Self

Table 4-4 Overview on retrieved web links [excerpt]

Finally, sentiment detection is implemented, particularly with regard to examine technology acceptance or gather opinions on technologies. Therefore, a rule-based classifier is applied and distinguishes between neutral, positive and negative tweets. Lists of positive respectively negative words are used to classify the tweets. Most tweets are neutral in the dataset on *#quantifiedself*, few are positive, and nearly none are negative (see Table 4-5 for examples). This implies that the discussions are more a neutral exchange of facts and information or technology hype but negligible on critical issues (e.g., ethics, privacy). This observation is consistent with the qualitative content analysis of the tweets.

Positive	**Neutral**	**Negative**
Anyone used this? It sounds like a fantastic app for some #QuantifiedSelf action.oggr: My Favorite Free Health App http://t.co/mOafyf5xlX	*Meetups This Week http://t.co/Nujsuk9jUC #quantifiedself*	*"The unexamined life is not worth living" ~ #Socrates #quantifiedself*
Nice little piece of hardware; pretty reliable so far #scanadu #health #quantifiedself http://t.co/6oFS7RGTix	*"What is a step anyway?" - @grapealope talking about #quantifiedself device accuracy at #SXSW*	*Exhaustion, obsession, self-tracking #quantifiedself http://t.co/QsykNvgA2i*

Table 4-5 Sentiment analysis: examples

As outlined by this use case, Twitter data analysis gives a first overview on a topic and captures the ongoing debate. Especially the societal discourse can be faster examined than by using classical methods as surveys or interviews and much more people (and thereby opinions) can be considered than otherwise. So the results deliver valuable input for the foresight exercise and support the process as such.

4.3.3 Twitter for User Engagement

In principal, Twitter is not only designed for one-way communication but can be used for user interaction to reach and involve the public in future planning processes. This can be transferred to foresight as illustrated next.

Process documentation

Since a while, Twitter has been used during events and conferences to announce talks or other advertisements. Moreover, Twitter might be used during larger workshops to gather additional input and ideas or give administrative information. Kelliher and Byrne (2015) use an online platform in their participatory foresight work for process documentation and additional discussions where Twitter was a small part of a larger framework. They see opportunities of their event-centered documentation in facilitating the dissemination of the content and the resulting community discourse. What is already common for conferences might be used during larger foresight workshops (more than 50 participants). Marked with an own workshop hashtag (e.g., the acronym of the project), additional comments can be given, instant feedback can be received, suggestions and discussion can take place, and the participants can stay informed after the workshop finished. Of course working only with Twitter offers less opportunities than a framework as in Kelliher and Byrne (2015). However, implementing Twitter is easier and it can be integrated in other online platforms.

Thinking about the future

For example, scenario development is a common method in foresight (e.g., van der Heijden, 2005; Reibnitz, 1991). Different factors and future developments are taken into consideration to illustrate possible futures. Therefore, different methods are applied, but in most cases the input is aggregated to short stories each describing one future. Platforms as Twitter can be used to gather relevant issues and ideas on how things might evolve (future paths). For example, Raford (2012) describes a pilot project in which he asked four questions on the future of public services using a specialized software platform. Participants from all over the world were invited to contribute brief stories or opinions tagged with keywords. Finally, the input was aggregated to three short scenarios. In a later article, Raford (2015) discusses the application of Web 2.0 in the context of scenario planning and criticizes the still limited use in foresight. He concludes that, principally, the advantages lay in (1) the huge sample of participants providing a greater data source and, thereby, delivering more input for the scenario process, (2) real-time monitoring of opinions and comparison to a common base, and (3) the rapid testing of scenario spaces in online communities.

Furthermore, Twitter can be used as alternative to survey tools. Instead of keywords as in Raford (2012), hashtags annotate the answers. Among those participating, the user data can be retrieved for sending a survey or getting interview contacts. In a first attempt, the following tweets were sent (see Figure 4-7). When initiating a discussion, the tweet should be marked with a unique hashtag that the respondents can tag their tweets (such as *#futureQS* or *#Quantfut*). By using relevant existing hashtags such as *#quantifiedself,* related communities are notified. Moreover, by adding key organizations or foundations (@message), the tweet can be spread further if the organization retweets the message (multiplicator function).

*How will **#quantifiedself** change our #future? Your #idea, #opinion, or **#vision** is requested! Be part of a future study #futureQS*

The **#future** of ***#quantifiedself***: What will it be like? What needs to be resolved? ***#Quantfut***

Figure 4-7 Tweets on the future of *quantified self*

Additionally, the Twitter community can deliver key points for drawing the scenario stories and Twitter is applicable for interactive story writing. For example, in 2012, The Guardian run an experiment with known authors on Twitter fiction writing (The Guardian, 2012). This might be linked to future visions. Guillo (2013) describes how he used an online survey embedded in an online platform for evaluating images of the future. He worked with young people (university students from Spain and Finland). One conclusion is that the participants missed interactive components. This may be resolved by Twitter where images of the future can be published online for discussion and receiving feedback. Further on, this relates to what Raford (2015) describes with the rapid feedback on scenario spaces by the online community. So visions can be tried to gather by initiating a discussion (see Figure 4-8).

> *What do you imagine the #future of **#quantifiedself** will be like? Come up with your **#vision** or story! **#QSfiction***

Figure 4-8 Future visions on *quantified self*

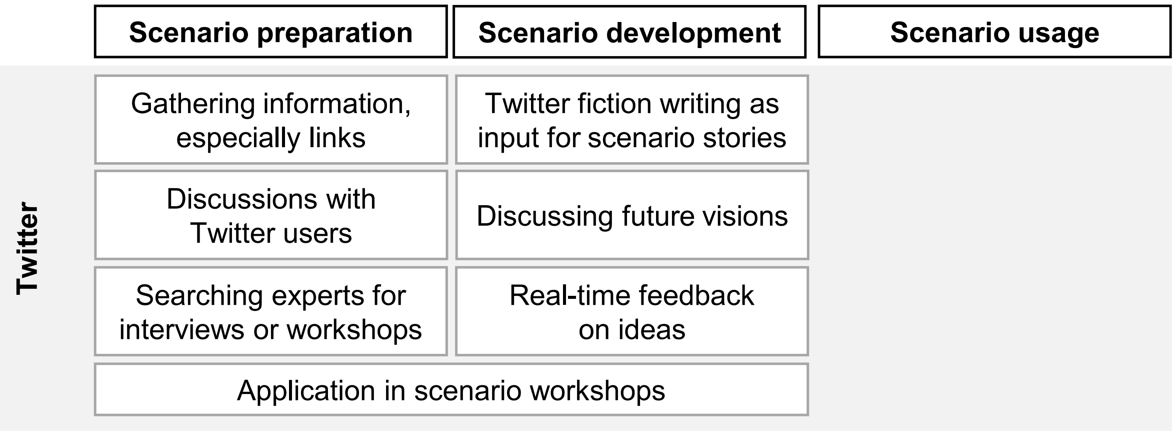
As a matter of fact, the second part on Twitter user engagement and interaction (Section 4.3.3) was more on showcasing the idea and illustrating possible realizations. The faint response emphasizes that more advertisement (over different channels) is necessary. Another important point is network building. It is central to be connected within the relevant communities or with many interested people for being recognized and getting attention. This of course needs some time and effort but increases the likelihood of success. Nevertheless, the variety and breadth of participants (e.g., age, social background) reached over Twitter cannot be reached over *classic* methods. Therefore, more effort and research should be conducted to get Twitter working for participatory foresight in future. Principally, it worked very well in other applications, see, e.g., on Twitter chats (Budak and Agrawal, 2013), for conducting surveys (Marwick and Boyd, 2011) or for *twitterviews* (interviews on Twitter).

4.4 Conclusion and Future Work

The following section summarizes the findings, discusses the limitations of Twitter for foresight and draws a conclusion, especially for future work with Twitter.

After a brief introduction to Twitter, the article describes the current state of research on Twitter. Based on this, first ideas on how to use Twitter for foresight were outlined. The research aim of this article was to examine the potential of Twitter for monitoring and interactive purposes in the context of foresight exercises. As the article showed, Twitter enables both monitoring and engagement. To quickly familiarize with a topic, the structures of Twitter but also data retrieval are most helpful. The content offers a broad picture reaching, for example, from science results, advertisement to news reporting and indicates *if* the topic gets any attention at all (number of tweets), *what* is discussed (e.g., used hashtags) and *how* emotional (sentiment analysis). Twitter facilitates a fast access to user-generated content, a huge number of people and real-time feedback on ideas. So, working with Twitter as communication and exchange platform broadens the perspective of foresight activities and enables the involvement of further stakeholders or different groups of the society not considered by foresight otherwise. Compared to other applications, no own interactive framework needs to be set up. This has, among others, the advantage that people are addressed in their known social media environments.

Summed up, a foresight exercise supported by Twitter may be designed as illustrated in Figure 4-9 on the example of scenario development. First, the initial step of information gathering may be facilitated by an automatic summary of the considered topic. Especially the web links contained in the tweets direct towards additional information, reports or websites and might deliver valuable input. In addition, expert search and discussions on Twitter may further enhance the preparation phase. As outlined above, workshop participants can contribute additional insights over Twitter. In the second step, when the input is processed to scenario stories, Twitter fiction writing might support the development of the stories and the discussion of future visions. Apart from, real-time feedback to questions and ideas from a global network can be received.



	Scenario preparation	Scenario development	Scenario usage
Twitter	Gathering information, especially links	Twitter fiction writing as input for scenario stories	
	Discussions with Twitter users	Discussing future visions	
	Searching experts for interviews or workshops	Real-time feedback on ideas	
	Application in scenario workshops		

Figure 4-9 Using Twitter in the context of scenario development

Of course, Twitter as an information source and exchange platform has limitations that should be kept in mind when designing applications. For instance, one drawback is that data cannot be retrieved retrospectively and each analysis starts from now on. But detecting trends would require data retrieval over a longer time. So, for the course of this article trend recognition was exclude and focused on other possible applications such as idea generation and monitoring, information exchange and participatory approaches as comprehensive stakeholder engagement. A further limitation is the *sender receiver*-fallacy: Tweets that are sent not necessarily read. Principally, Twitter as a data source is accused for not being representative (see, e.g., Bruns and Stieglitz, 2014), to be no reliable scientific source and generally a product of marginal groups. Of course Twitter analysis is not representative for society (see for a discussion e.g., Boyd and Crawford, 2012), but it reflects ongoing developments and changes. In addition, information credibility is an issue on Twitter (e.g., Castillo et al., 2011). Nevertheless, when actively using Twitter as a supportive element in foresight as, e.g., in scenario development, biased user activity or representativity of the data are less critical because foresight asks for possible, desirable, or provocative futures or statements regarding future developments. So *correctness* is less an issue than diversity and heterogeneity of the received input as, for example, from Twitter.

Summarized, Twitter has limitations as every method or data source. So the observations from Twitter should be weighed up against other data and foresight activities or be part of a larger framework. Therefore, a mixed methods approach should be applied instead of building only on Twitter analysis. Apart from the example of scenario development above, of course, many other possible applications combining Twitter and foresight methods exist. This relates, for example, to real-time Delphi or new foresight gaming approaches, but also to the examination of organizational networks, technology acceptance, or the comparison to other frequently used data sources (e.g., patents, scientific publications). Still, this also relates to the limited use of Web 2.0 applications for foresight in general. Hence, many options for future work related to methodological integration or combination evolve but also the applied context (e.g., policy planning, corporate strategy development). So the added value needs to be individually examined. Moreover the social structure and the knowledge structures of Twitter enable new research opportunities while the understanding of the structures is most relevant for the interpretation of the findings (see, e.g., Jungherr, 2015). Finally, Twitter has offerings to the foresight community such as networking among researchers, the spread of information (invitations to conferences, workshops, project results, etc.), or triggering online discussions on certain issues.

As described above, the intention of this article was to gather first ideas and create a starting point for detailed future research. However, this article points out a wide range of promising opportunities how to use Twitter in future foresight applications and it remains to be seen which will get realized.

References

Amanatidou, Effie; Butter, Maurits; Carabias, Vicente; Konnola, Totti, et al. “On concepts and methods in horizon scanning: Lessons from initiating policy dialogues on emerging issues.” *Science and Public Policy* 39, no. 2 (2012): 208–221.

Asur, Sitaram; Huberman, Bernando A. “Predicting the Future with Social Media.” International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM, 2010: 492–499.

Boyd, Danah; Crawford, Kate. “Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon.” *Information, communication & society* 15, no. 5 (2012): 662–679.

Boyd, Danah; Golder, Scott; Lotan, Gilad. “Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter.” 43rd Hawaii International Conference on System Sciences (HICSS). Kauai, Hawaii, 2010.

Bruns, Axel; Burgess, Jean. “Notes towards the Scientific Study of Public Communication on Twitter.” In *Science and the Internet,* edited by Alexander Tokar, Michael Beurskens, Susanne Keuneke, Merja Mahrt, Isabella Peters, Cornelius Puschmann, T. van Treeck and Kathrin Weller. Düsseldorf: düsseldorf university press, 2012: 159–169.

Bruns, Axel; Stieglitz, Stefan. “Twitter data: What do they represent?” *it - Information Technology* 56, no. 5 (2014): 240–245.

Budak, Ceren; Agrawal, Rakesh. “On participation in group chats on twitter.” edited by Daniel Schwabe. Proceedings of the 22nd international conference on World Wide Web companion. Rio de Janeiro, Brazil, 2013: 165–176.

Cachia, Romina; Compañó, Ramón; Da Costa, Olivier. “Grasping the potential of online social networks for foresight.” *Technological Forecasting and Social Change* 74, no. 8 (2007): 1179–1203.

Castillo, Carlos; Mendoza, Marcelo; Poblete, Barbara. “Information credibility on twitter.” Proceedings of the 20th international conference on World wide web. Hyderabad, India, 2011: 675–684.

Gaffney, Devin; Puschmann, Cornelius. “Data Collection on Twitter.” In *Twitter and society,* edited by Katrin Weller, Axel Bruns, Jean Burgess, Merja Mahrt and Cornelius Puschmann, 2014: 55–67.

Gheorghiou, Radu; Curaj, Adrian; Paunica, Mihai; Holeab, Cosmin. “Web 2.0 and the Emergence of Future oriented Communities.” *Economic Computation & Economic Cybernetics Studies & Research* 43, no. 2 (2009): 1–11.

Globalwebindex. “GWI Social Summary Q4 2014.” 2014. http://www.globalwebindex.net/blog/instagram-still-lags-twitter-as-the-fifth-biggest-social-network.

Grubmüller, Verena; Götsch, Katharina; Krieger, Bernhard. “Social media analytics for future oriented policy making.” *European Journal of Futures Research* 1:20 (2013).

Guillo, Mario. “Futures, communication and social innovation: using participatory foresight and social media platforms as tools for evaluating images of the future among young people.” *European Journal of Futures Research* 1, no. 1 (2013): 1–17.

Haegeman, Karel; Cagnin, Cristiano; Könnölä, Totti; Dimitrov, Georgi; Collins, Doug. “Web 2.0 foresight for innovation policy: A case of strategic agenda setting in European innovation.” *Innovation* 14, no. 3 (2012): 446–466.

van der Heijden, Kees. *Scenarios: The art of strategic conversation*. 2nd ed. Chichester, West Sussex, Hoboken, N.J.: John Wiley & Sons, 2005.

Holmberg, Kim; Thelwall, Mike. “Disciplinary differences in Twitter scholarly communication.” *Scientometrics* 101, no. 2 (2014): 1027–1042.

Java, Akshay; Song, Xiaodan; Finin, Tim; Tseng, Belle. “Why we twitter: understanding microblogging usage and communities.” Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis. San Jose, CA, USA: ACM, 2007: 56–65.

Jungherr, Andreas. *Analyzing Political Communication with Digital Trace Data: The Role of Twitter Messages in Social Science Research*. Cham, s.l.: Springer International Publishing, 2015.

Kelliher, Aisling; Byrne, Daragh. “Design futures in action: Documenting experiential futures for participatory audiences.” *Futures* 70 (2015): 36–47.

Marwick, Alice E; Boyd, Danah. “I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience.” *New media & society* 13, no. 1 (2011): 1–20.

Metaxas, Panagiotis Takis; Mustafaraj, Eni; Wong, Kily; Zeng, Laura, et al. “Do Retweets indicate Interest, Trust, Agreement?” *CoRR* arXiv preprint arXiv:1411.3555 (2014).

Pang, Alex S.-K. “Social scanning: Improving futures through Web 2.0; or, finally a use for twitter.” *Global Mindset Change* 42, no. 10 (2010): 1222–1230.

Priem, Jason; Piwowar, Heather A; Hemminger, Bradley M. “Altmetrics in the wild: Using social media to explore scholarly impact.” *CoRR* abs/1203.4745 (2012).

Puschmann, Cornelius; Burgess, Jean. “The Politics of Twitter Data.” In *Twitter and society,* edited by Katrin Weller, Axel Bruns, Jean Burgess, Merja Mahrt and Cornelius Puschmann, 2014: 43–54.

Raford, Noah. “Crowd-sourced Collective Intelligence Platforms for Participatory Scenarios and Foresight.” *Journal of Futures Studies* 17, no. 1 (2012): 117–128.

Raford, Noah. “Online foresight platforms: Evidence for their impact on scenario planning & strategic foresight.” *Technological Forecasting and Social Change* 97 (2015): 65–76.

Reibnitz, Ute. *Szenario-Technik: Instrumente für die unternehmerische und persönliche Erfolgsplanung*. Wiesbaden: Gabler, 1991.

Risse, Thomas; Peters, Wim; Senellart, Pierre; Maynard, Diana. “Documenting Contemporary Society by Preserving Relevant Information from Twitter.” In *Twitter and society,* edited by Katrin Weller, Axel Bruns, Jean Burgess, Merja Mahrt and Cornelius Puschmann, 2014: 207–219.

Russell, Matthew A. *21 Recipes for Mining Twitter*. Sebastopol: O’Reilly Media, Inc; O'Reilly Media Inc, 2011.

Schatzmann, Jörg; Schäfer, René; Eichelbaum, Frederik. “Foresight 2.0 - Definition, overview & evaluation.” *European Journal of Futures Research* 1, no. 1 (2013): 1–15.

Stieglitz, Stefan; Krüger, Nina. “Public Enterprise-Related Communication and its Impact on Social Media Issue Management.” In *Twitter and society,* edited by Katrin Weller, Axel Bruns, Jean Burgess, Merja Mahrt and Cornelius Puschmann, 2014: 281–292.

Swan, Melanie. “Sensor Mania! The Internet of Things, Wearable Computing, Objective Metrics, and the Quantified Self 2.0.” *Journal of Sensor and Actuator Networks* 1, no. 3 (2012): 217–253.

Teevan, Jaime; Ramage, Daniel; Morris, Merredith R. “#TwitterSearch: A comparison of Microblog Search and Web Search.” Proceedings of the fourth ACM international conference on Web search and data mining. Hong Kong, China: ACM, 2011: 35–45.

Terpstra, Teun; de Vries, A; Stronkman, R; Paradies, G. L. “Towards a realtime Twitter analysis during crises for operational crisis management.” ISCRAM 2012 conference proceedings book of papers. Vancouver, Canada: Simon Fraser University, 2012: 1–9.

The Guardian. “Twitter fiction: your 140 character stories.” *The Guardian,* October 15, 2012.

Thelwall, Mike; Haustein, Stefanie; Larivière, Vincent; Sugimoto, Cassidy R. “Do Altmetrics Work? Twitter and Ten Other Social Web Services.” *PloS one* 8, no. 5 (2013): e64841.

Twitter. “About Twitter.” 2015. https://about.twitter.com/company, accessed April 2015.

5 Web-based Scenario Development: Process Improvements

Abstract: Scenario development is an established foresight method. However, scenario processes require much time, and most methodological developments concentrate on the specific task of developing consistent scenarios. One of the key challenges in scenario development is to capture the topic and identify its key influences. This has potential for improvements. In times of big data, far more options exist for the rapid exploration of a topic than manual literature analysis. Hence, this article examines web and text mining for their usability for data retrieval and aggregation in the context of scenario development. To better present the argument about the benefit, scenario stories are formulated. In this article, a new scenario process is proposed and described, using the future of the topic *quantified self* as an example. As the results show, web and text mining deliver a very good starting point for discussing the scenario content. The rapid overview with the visualizations remarkably reduces the reading effort. Still, future projections need to be searched manually, but the results from the automatic analysis comprehensively guide this step.

Keywords: Web Mining, Foresight, Text Mining, Twitter, Scenario Development, Topic Modeling, Social Media, Morphology Analysis

5.1 Introduction

Scenario development is a popular and established method, especially in foresight, to depict different futures (van der Heijden, 2005). One crucial challenge at the beginning of scenario development is to capture the topic and identify its key influences. However, technical and methodological improvements of scenario development, in most cases, concentrate on later stages of the process (e.g., Amer et al., 2013). Still, the process of scenario development is time-consuming and process improvements might increase its applicability. As practical experience shows, three points in the process of scenario development require much time: desk research (e.g., Kuosa, 2012), literature analysis (e.g., Mietzner and Reger, 2005) and combination of future alternatives for different scenarios (as e.g., by consistency analysis or cross-impact analysis). Scenario processes have often been accused of being too qualitative (Hirsch et al., 2013). Moreover, the need for a combination of qualitative and quantitative methods is frequently emphasized for foresight in general (Haegeman et al., 2013). But the combination of qualitative and quantitative approaches is seen as a methodological challenge (van Notten et al., 2003). As a further point, recent foresight and its methods still do not profit from the epistemic value of *big data*, although this holds potential to improve foresight by making sensible use of new quantitative tools. For example, social media is widespread, but foresight rarely examines the evolving opportunities.

Building on these critiques, this paper examines the potential of Twitter as a data source for future scenarios. Twitter, as a widely used social media platform, covers a broad spectrum of content (see Section 4). Apart from the hashtags, especially the web links contained in the tweets referring to further content will be explored here. Further on, the use of text mining (Feldman and Sanger, 2008) to aggregate information is described. This fosters a rapid overview of aspects describing the scenario field to capture the topic and derive influence areas and factors (Kayser and Shala, 2014). To better argue the contribution of this form of data analysis, this article will not only concentrate on the preparation of the scenario field. This article will also consider the continuing process and develop scenario stories based on the results from web and text mining.

In summary, the aim of this article is to show a new way of optimizing the process of scenario development in order to have more time for the final tasks of foresight: formulate recommendations, enforce decisions, develop a future strategy and initialize an action plan. Therefore, a new method is developed, building on web and text mining. This is illustrated for the future of *quantified self*.

The article begins with an introduction to scenario development in Section 5.2. Next, the methodology for web-based scenario development is introduced in Section 5.3, using *quantified self* as an example. The results are discussed and final conclusions are drawn in Section 5.4.

5.2 Improving Scenario Development

Scenarios illustrate different possible futures. As an established foresight method, scenarios serve as a framework to think about possible future developments in order to derive robust strategies (van der Heijden, 2005; Reibnitz, 1991). Many different scenario approaches exist, and the scenario processes can be aggregated into three steps (see, e.g., O'Brien and Meadows, 2013). After setting the scope and purpose of the project, information is needed on the subject of interest in the first step. Hence, a deep knowledge and understanding of the scenario field are most relevant for the success of the process (Kuosa, 2012; Mietzner and Reger, 2005). The insights are aggregated to influence factors (e.g., *market, privacy*). Next, future alternatives are formulated for each factor. In the second step, the interdependencies between the alternatives are analyzed, in order to draw consistent future scenarios. In the third step, the scenarios are applied in areas such as strategy development (Godet, 1997), foresight (Bezold, 2010), or technology assessment (Grunwald, 2010).

As practical experience shows, three points in particular in the process of scenario development require much time and effort: desk research, literature analysis and the combination of the future alternatives to different scenarios (Mietzner and Reger, 2005; Kuosa, 2012; Raford, 2015). The following discusses improvements for these three challenges, as indicated in Figure 5-1.

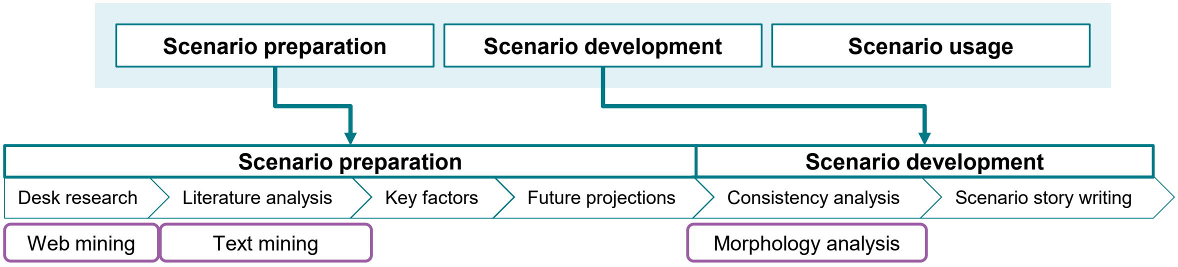


Figure 5-1 Process of scenario development

5.2.1 Data Gathering

This article considers the use of web data for scenario development. Generally, the two aspects of collecting data for scenarios are participatory approaches (e.g., workshops, interviews, focus groups) or desk research (see, e.g., van Notten et al., 2003).

In terms of participation and social media, Raford (2012) configures an online platform to discuss the future of public services. In his case study, a global set of respondents was able to send stories or opinions tagged with keywords. Finally, three scenarios are formulated. In a later article, Raford (2015) discusses the application of Web 2.0 in the context of future scenarios, and criticizes the fact that the use in foresight is still limited. His conclusions concerning the use of social media for scenario development emphasize the benefit of the huge sample of participants, the real-time character and the rapid feedback. However, the content and (strategic) alignment of scenarios is something internal and not publicly discussed in many cases, while the huge audience is, of course, an interesting aspect. In contrast, interviews or workshops restrict the received input to a small number of people. Another critique on participatory approaches for data collection are mental models (van der Heijden, 2005) and the reliance on expert statements per se.

This article will concentrate on improving desk research and systematic data collection. The method developed in this article will be applicable to broad *societal* issues. Therefore, a focus on scientific developments, as bibliometric data implies (see, e.g., Stelzer et al., 2015), is not an aim of this article, but to examine the contribution of web content for systematic data gathering.

No related work has tried to use web content, based on automatic retrieval, in the context of future scenarios so far. However, one challenge is to find a set of appropriate websites related to a certain topic. This implies a high search effort. Therefore, the process starts with retrieving data from Twitter. In general, Twitter, a micro-blogging service established in 2006, is nowadays both a social network and information sharing platform (Java et al., 2007; Kwak et al., 2010). Twitter has a broad spectrum of content and contributors. Originally, Twitter was more a platform for merely private exchange, but has evolved into a network that attracts professional interests, such as enterprises monitoring customer interests and opinions (Bruns and Burgess, 2012). The basic element of Twitter are tweets as short messages, many of which contain web links referring to external content, such as blog entries or news articles. Twitter is used as an information base for future scenarios because data can easily be retrieved using its API by searching for hashtags or terms. This reduces the search effort and desk research can be largely automated. Second, Twitter aggregates content from various other platforms, such as newspapers or blogs. So, the diverse set of contributing actors (such as private persons, companies, or organizations) enables to capture different perspectives from only one platform. Thereby, this data contains a broad spectrum of stakeholders and views. Third, as concluded in Section 4, the web links, in particular, are compelling to deliver information for foresight exercises and might describe the topic very well. This will be further explored in this article.

5.2.2 Information Aggregation

Usually, preparation of scenarios entails a high reading effort (see, e.g., Mietzner and Reger, 2005). This is distributed among a set of people, thereby getting a natural bias, caused by different personal interests and foci. However, the scenario preparation might be improved by text mining (see, e.g., Feldman and Sanger, 2008) to reduce this reading effort, capture the topic and deliver a common base for discussing the influence factors (Kayser and Shala, 2014). The following introduces two approaches for information aggregation based on text mining: concept mapping and topic modeling.

Concept mapping

Concept mapping aggregates texts to concepts, based on the words they contain (Leximancer, 2011; Stockwell et al., 2009). In previous work, concept mapping has been applied on different datasets. For example, concept mapping is applied on Twitter data to distinguish nutrition patterns (Vidal et al., 2015) or used to map blogs and tweets about social media (Cameron et al., 2011). Bell and Seidel (2012) analyze interview transcripts, while Angus et al. (2013) examine a conversation transcript and the contribution of each agent. Cretchley et al. (2010) explore communication strategies of carers interacting with people with schizophrenia. Others analyze scientific articles (Liesch et al., 2011; Poser et al., 2012) or their abstracts (Anagnostopoulos and Bason, 2015; Rooney et al., 2011). Davies et al. (2006) analyze the textual comments of a survey. They summarize the benefit of *Leximancer* as "*[…] a useful tool when a researcher is exploring the textual data to attempt to uncover important factors. In other words, it is highly useful when the researcher does not have an a priori set of factors or model by which to analyze the data* (Davies et al., 2006)." Building on this experience, concept mapping is transposed to scenario development to summarize the topic and identify influence factors.

In technical terms, concept mapping applies naïve Bayes classification (Yarowsky, 1995; Salton, 1988), and operates in two steps. First, a classifier is constructed (semantic extraction). Thereafter, a categorical coding scheme is learned. Stopwords are then removed, word variants are merged, and the frequency of single words and their co-occurrence are calculated. A concept as a group of related terms is built by a thesaurus as a term classifier. The concept bootstrapping algorithm results in concept seeds as start values for the concepts as clusters. The second step corresponds to *coding,* as in content analysis. Here, text segments of one up to three sentences are classified. Additionally, relations are identified within and between the concepts. Finally, the concepts are denoted as nodes aggregated to themes, which are clusters of concepts and illustrated as bubbles. The algorithm of concept mapping is objective and repeatable (see Smith and Humphreys, 2006 for an evaluation), and might overcome the natural bias of manually searching factors from texts. In addition, relations between the concepts (distance on the map) and the relevance of the theme (size) are indicated. Furthermore, the theme size can be varied.

Topic modeling

This second approach uses topic modeling and thereby uses an algorithm that is different from the one used in *Leximancer*. This delivers an additional perspective on the data. One critique on *Leximancer* relates to its term extraction mechanism and its difficulties in extracting nouns. Therefore, *PoS-tagging* is implemented here (Bird et al., 2009). The automatic identification of the grammatical instance facilitates the extraction of chains of nouns, such as *heart rate* or *apple watch*. Therefore, the text is cut into words and the grammar is examined sentence-wise.

For analyzing the data, topic modeling is applied, using *Latent Dirichlet Allocation* (Blei et al., 2003). For example, topic models are applied to analyze topics in publication data (Yau et al., 2014). This unsupervised approach has first been applied in the context of scenario development in Kayser and Shala (2014). Topic models are useful in structuring texts when there is no domain knowledge of the subject covered *(Blei and Lafferty, 2007).* Topic models reveal the hidden thematic structure in texts, while topics relate to influence areas and factors. A statistical model is inferred during the generation process, and soft clustering is applied (words might belong to more than one topic) (Miner, 2012). The underlying assumption is that documents are built from topics drawing words from word distributions. A topic is distributed over a fixed vocabulary. The topics are denoted by a probability mass function over each possible word. Topics have associated term probabilities and for each document, topic proportions are computed (likelihood of a topic to appear in a document). For this article, the *gensim*-package was implemented (Řehůřek and Sojka, 2010). For a better performance, the stream of values is split into smaller chunks (500 words). The final set of topics is manually labeled. This second approach is implemented in Python and SQL.

5.2.3 Into the Future

After the state-of-the-art is summarized, there remains the question of how to proceed into the future. At this point, explorative scenario approaches are eligible, based on influence factors (e.g., Börjeson et al., 2006; Glenn and The Futures Group International, 2009; Kosow and Gassner, 2008). Therefore, different assumptions are formulated as to how specific factors may evolve in future. These are combined into logical, plausible and consistent scenarios. In this process, all possible combinations of future projections are considered. Therefore, different approaches exist (see for an overview, e.g., Kosow and Gassner, 2008; Bradfield et al., 2005). Predominantly, consistency analysis is used at this point (see, e.g., Gausemeier et al., 1996). However, critique on consistency analysis relates to the time effort and the focus on consistent futures. First, discussing the consistency matrix needs much time and the main point is that projections match or contradict each other. Second, calculating consistency values highlights consistent scenarios. But in scenario processes, the focus equally lies on discussing different futures. Therefore, the consistency value is an inappropriate measure, and does not indicate differences between the scenarios.

Because of these reasons, morphology analysis (Zwicky, 1948) is used in this article. First, the system as the scenario space is described and broken into its single components (factors and projections). Next, the projections are systematically combined in a morphological space, enforcing consistent scenarios. Thereby, exclusions and preferences among the different projections are identified (e.g., Godet, 1997). So, each selection influences the number of possible combinations, implying path dependencies. Finally, scenarios are formulated from the combination of projections.

The advantages of morphology analysis compared to consistency analysis or cross-impact analysis lie in the fewer number of steps that have to be taken to compose scenarios. First, it is argued which projections match or do not match. As the next step, plausible stories are drawn without calculating consistency values. Starting at one projection highlights where to continue the story line (exclusive and preferred links). From the beginning, the focus lies on getting different scenarios, rather than only consistent scenarios (as for consistency approaches). Morphology analysis helps eliminate contradictions and analyze different combinations of factors in a graphical representation to ensure plausibility (e.g., Amer et al., 2013).

5.3 Methodology: Web-based Scenario Development

This section introduces the methodology for web-based scenario development (see Figure 5-2). This explorative scenario approach begins by collecting data based on a Twitter search. This data is aggregated to retrieve factors (tweet analysis). First, the hashtags contained in the tweets are analyzed. Second, links are extracted from the tweets, and web mining is applied to retrieve the text of the websites. From this, more in-depth insights are expected. Following this, concept mapping and topic modeling are applied on the texts retrieved from the websites. Based on these results, factors are derived and discussed in the scenario team and future projections are formulated for the resulting list of factors. Finally, morphology analysis is conducted to develop the scenario stories.

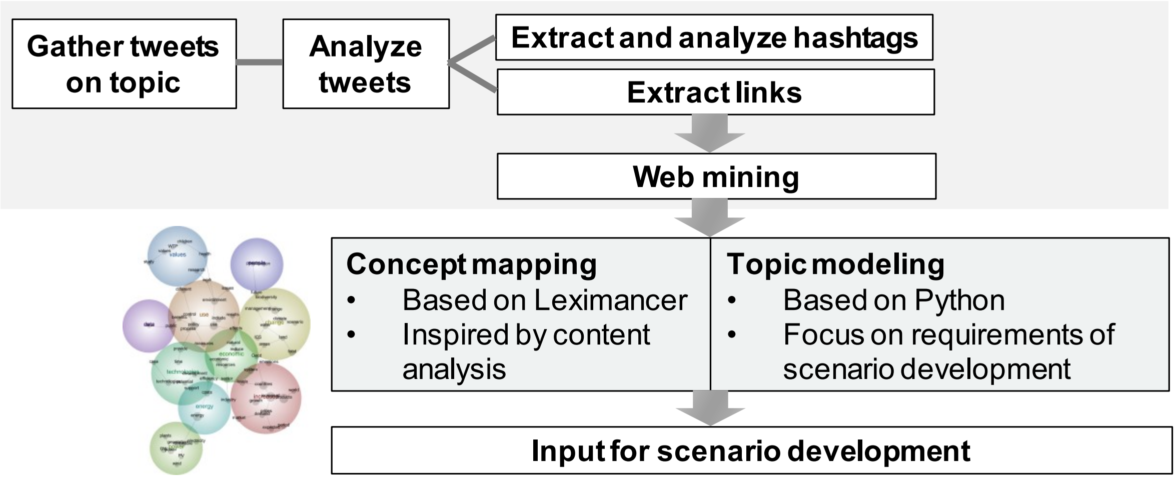

Gather tweets on topic
Analyze tweets
Extract and analyze hashtags
Extract links
Web mining
Concept mapping
• Based on Leximancer
• Inspired by content analysis
Topic modeling
• Based on Python
• Focus on requirements of scenario development
Input for scenario development

Figure 5-2 Process of web-based scenario development

In the following, the method is explained with reference to the example of *quantified self*. *Quantified self* describes self-monitoring and self-tracking applications to monitor (physiological) variables, such as heart rate, blood pressure or eating habits (see, e.g., Swan, 2012). In the case of this article, a time-frame of five years was chosen as the exemplary planning horizon.

5.3.1 Retrieving Data from Twitter

Data is retrieved using the Twitter API. In the case of this article, this was done for tweets containing the hashtag *#quantifiedself.* Table 5-1 summarizes the key parameters of the dataset. The data covers half a year. The original number of 24.850 tweets was cleaned and 6.614 English tweets were further processed. As described in the following, the hashtags and the contained links were analyzed.

Search for	*#quantifiedself*
Time	*Tweets from 2 May 2015 to 2 September 2015*
Number of tweets in total	24.850 (15.776 retweets)
Number of English tweets	18.433
Number of English tweets (retweets excluded)	6.614

Table 5-1 Key parameters of the dataset

5.3.2 Analyzing the Hashtags

The analysis of the hashtags gives a first impression of the ongoing discussion. The hashtags are mapped as a network. The size of the node relates to how often a hashtag is mentioned with other hashtags, while stronger links imply stronger ties.

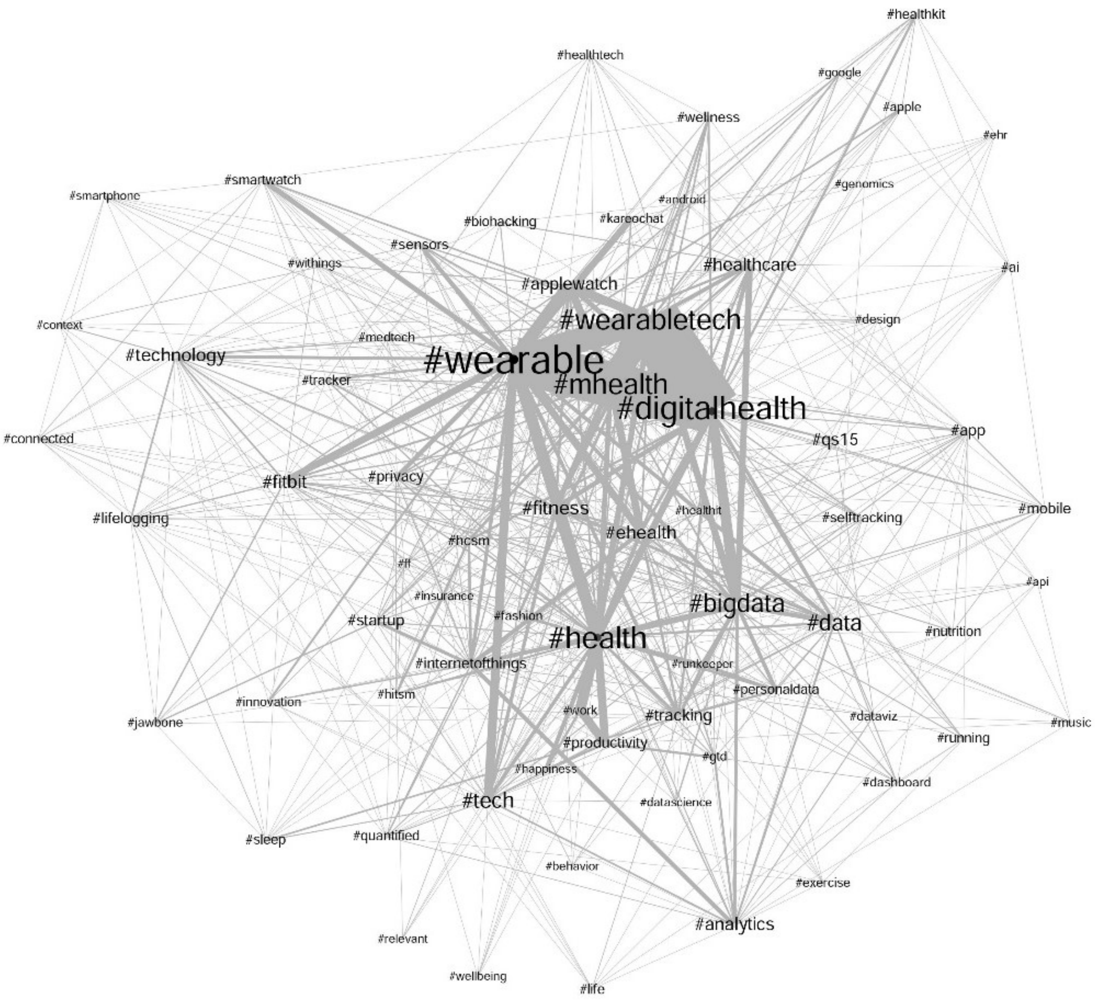


Figure 5-3 Hashtag network (node degree ≥ 10)

As Figure 5-3 illustrates, the three most central hashtags are *#wearable*, *#mhealth, #digitalhealth* and *#wearabletech*. Regarding thematic clusters, there is one on data in the bottom right corner (e.g., *#bigdata*, *#data*, *#analytics*). In the center of the network, there are entries about health (e.g., *#health*, *#digitalhealth*, *#healthcare).* And wearables and devices are frequently mentioned (e.g., *#wearable*, *#wearabletech*). However, *#privacy* and *#fitness* are also covered in this network.

5.3.3 Web Mining

As discussed in Section 2.1, many tweets contain web links referring to additional content. Therefore, web mining (see, e.g., Liu, 2011) is applied to retrieve the websites underlying these web links. Thereby, more content can be processed than by manual desk research, and the websites can later be processed by text mining.

For web mining, the Python package *beautifulsoup* was used. First, duplicate web links were eliminated. In addition, web links that obviously direct to images, very short texts or videos were removed (such as *YouTube*, *Instagram* or *Vimeo*). Owing to the fact that English tweets do not necessarily direct to English websites, a language check was conducted on the header of the website to exclude non-English content from the following analysis. Next, the text on the websites was retrieved and sections marked with a <p>-tag were stored. Websites containing less than 500 characters were not stored (as e.g., advertisements). Finally, 1.322 websites were retrieved and stored. After a manual check, further cleaning led to a final dataset of 1.318 texts. Table 5-2 gives an overview on the five most frequent websites from which content was retrieved. Among these 519 different websites in total, are blogs, such as *dacadoo*, but also news channels, such as *wired or mobihealthnews*. This underlines the spread and variety of the content used for the further analysis.

Websites	Number of retrievals
quantifiedself.com	113
mobihealthnews.com	61
meetup.com	46
medium.com	39
exist.io	28
engadget.com	26
wired.com	23
quantselflafont.com	21
linkedin.com	18
blog.dacadoo.com	18

Table 5-2 Top 10 websites from which data was retrieved

5.3.4 Aggregating the Content

In the next step, the content is aggregated to identify factors based on the retrieved websites. Therefore, text mining is applied. This is expected to be faster than reading through all the websites. For identifying influence areas and factors, the text from the websites runs through the two different approaches introduced in Section 5.2.2, concept mapping and topic modeling.

Concept mapping

The retrieved websites are read in *Leximancer* and are automatically processed. In this case, the standard settings were used (Leximancer, 2011), but with an adapted stopword list, merged word variants and the initial set of concepts were adapted. The concept map is denoted in Figure 5-4 and reveals eight different themes (*data*, *users*, *social*, *work*, *people*, *experience*, *rate*, *Apple watch*). Heart rate and sleep monitoring build an own theme and are closely linked to the *data*-theme, containing concepts such as *tracking*, *movement* and *activity*. In addition, the *data*-theme covers two important applications of quantified self: health and fitness. Health is at the center of this theme (*health*, *medical*, *care*, *patients*). While the *Apple watch* makes up an own theme, the technical components are included in the *data*-theme, such as *wearable*, *app*, *devices*, and *technology*. The *users*-theme is adjacent, containing *market*, *company* and *research* on the intersection with the data-theme. The market aspect of quantified self has not been discussed so far, but additional desk research showed that huge revenues are expected (Business Insider, 2013). The *experience*-theme and the *social*-theme both cover social media aspects and issues such as *privacy* or *control*. One theme is related to *work,* indicating that quantified self is increasingly used in professional environments, and may have an impact on the future of work. The *people*-theme highlights the impact quantified self can have on daily life by concepts such as *change* or *study*.

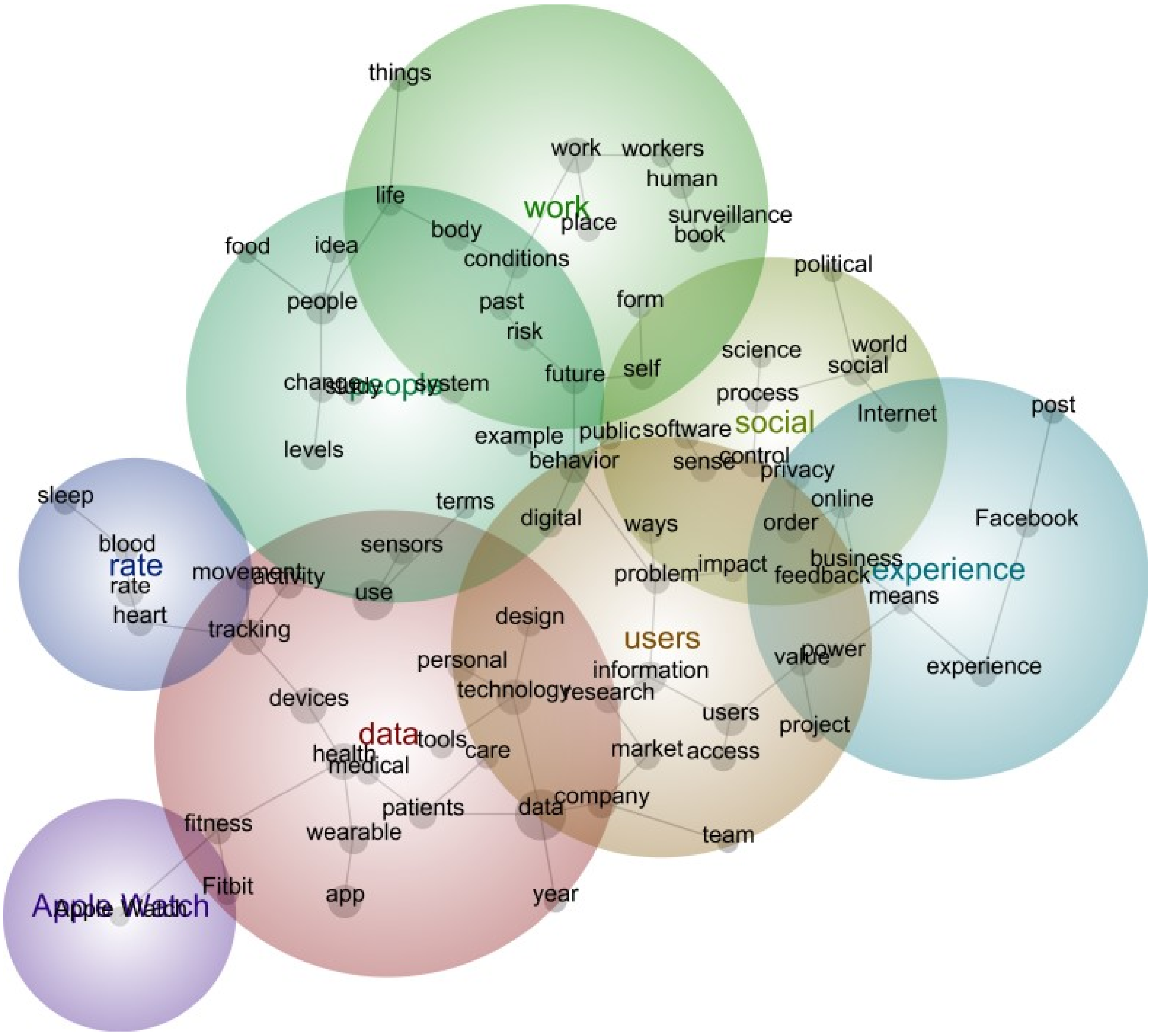


Figure 5-4 Concept map (theme size: 51%)

Topic modeling

The results of topic modeling deliver a further perspective on the texts and are denoted in Table 5-3. Terms occurring at least 30 times in the whole dataset were considered in this analysis. An iterative process showed that topic modeling showed the best results for five topics. The first topic describes the general potential of quantified self for applications such as health records. The second topic contains wearables and devices from different suppliers, such as *apple* and *fitbit*. The third topic indicates much exchange and networking on *quantified self,* due to terms such as *meetup*, *conference*, or *group*. This aspect did not show up so clearly in the previous analysis, but indicates a lively debate between the users and indicates a market for *quantified self*. The fourth topic is on data and related technologies. Here again, health-related aspects are closely connected (*health*, *patient*). Finally, the fifth topic is on monitoring and tracking. For example, this relates to measuring sleep patterns or activity profiles in general. However, the results of topic modeling underline that data, health, and wearables are topics that should be considered in the scenario process. Further points to be considered are the market potential of quantified self and potential user concerns with the monitoring and tracking applications.

Topics	
1 – Potential of quantified self	potential, search, condition, health record, secret, theme, charge, decision, code, exist
2 – Wearables & devices	apple watch, fitbit, watch, fitness tracker, exist, device, Melbourne, apple, smartwatch, band
3 – User exchange	Meetup, toolmaker, talk, event, funding, show, program, conference, group, check
4 – Data & technologies	Data, device, technology, people, patient, health, wearable, life, thing, company
5 – Monitoring & tracking	Sleep, monitor, length, memory, withings, human, activity tracker, role, quality, jawbone

Table 5-3 Topic model

5.3.5 Influence Factors

To better present the argument about the advantage of web and text mining, scenario stories are developed in the following. To begin with, the results from text mining were discussed in the scenario team to formulate influence factors. Additional desk research was conducted for a more detailed view to supplement and validate the results and obtain additional facts and statistics. This is relevant for arguing the future projections as the next step, and was supported by the pre-structuring from the results of text mining.

Finally, this leads to six influence factors. The basic technology underlying most quantified self-applications are wearables and devices (Factor 1), for example, to record vital signs. The whole topic has two main areas of application: sports & fitness (Factor 2) and healthcare (Factor 3). Statistics are important in sports anyway, and by using quantified self-devices, one can virtually compete with anyone by using social media. The digital health industry is increasingly making use of interconnected technologies to improve care quality and early detection. Moreover, data and its analysis (Factor 4) are relevant, and this gives user-generated content a new dimension. The recorded data (such as calorie intake, heartbeat, etc.) provides many opportunities for data analytics; although this data is valuable, it is also most sensible. This, of course, leads to privacy issues and user concerns (Factor 5). Autonomy, lifelogging and self-tracking are aspects covered under this factor. In addition, this relates to law and regulatory issues. Finally, market and business opportunities (Factor 6) evolve. Insurance companies are especially developing new business models at the moment. Further on, new industries are arising around health economics. As the concept map indicated, another aspect is the quantified workplace.

5.3.6 Future Projections

Future projections describe different ways of how things might evolve. However, this step cannot be automated due to the fact that the formulation of future projections requires detailed knowledge (e.g., statistics, facts, and numbers). So the required information is on a more detailed level than the results from text mining, and further investigation is appropriate. Principally, expert workshops are a useful tool for this step. As these are time-consuming with regard to the effect and output, this step was replaced by a structured desk research in order to make extensive and in-depth use of the already conducted web-based analyses. Apart from the visualizations from text mining, the database of the stored websites can be screened, hinting at relevant aspects. This finally leads to the future projections, as described in Table 5-4.

Influence Factors		Future Projections
(1) Healthcare	A	New applications for the early detection of diseases are evolving. Healthcare apps are widely spread together with established reward systems for effective use of healthcare wearables and apps (e.g., Siegel Bernhard, 2015).
	B	Health monitoring has reached a new dimension. For example, epigenome data analysis is an established technique (e.g., GIbbs, 2014). The data is used in healthcare for therapies and prevention of cancer, diabetes and obesity.
	C	Wearables are used just for fun. There are few working health applications.
(2) Fitness & Sports	A	It is very common to be “wired” when engaging in sports, and self-optimization is a key principle (see, e.g., Weintraub, 2013). As a side effect, insurances pay incentives for sports with fitness duties to be fulfilled on a regular basis.
	B	Devices in sports are not common, unless in professional sports.
(3) Wearables & Devices	A	Anyway, the hype is over. Some buy bracelets, but nobody wears them for more than half a year, as is already noticed today (Arthur, 2014). So, there is a decreasing number of new applications.
	B	There are a high spread and acceptance. Many devices include sensors and interfaces. In addition, the devices get cheaper and are mass-produced. Smart clothes are trending (Gibbs, 2014) and the sensors are invisible (Gartner, 2014).
(4) Data Analytics	A	Data is analyzed and increasingly used. Many apps and analysis platforms are developed. Furthermore, data is recombined to make sense of it, such as better understanding side effects of drugs.
	B	The many sensors and devices lead to an information overload; but nobody knows what to do with this data. Problems especially evolve at the interfaces, due to many small suppliers and decentralized storage.
(5) Privacy & Data Sharing	A	Quantified self has very high acceptance rates. It is normal that every movement and body function is tracked. For example, sleep patterns are used at court (see, e.g., Olson, 2014). Principally, consumers are willing to share their data, e.g., in reward to a reduced insurances (PWC, 2014).
	B	Users are afraid that wearables invade their privacy. Quantified self is rejected and many people feel reluctant to be tracked. Increasingly, leaks in data management and data misuse cause a further reduction of their acceptance.
	C	Consumers are becoming increasingly concerned about wearables and privacy. While 59% were concerned in 2014 (PWC, 2014), this number has increased in the meantime.
(6) Market & Business Opportunities	A	New business models and offers evolve, e.g., for insurances (Beuth, 2015) or targeted advertisements (Tozer, 2015). People with overweight and low performance rates have increasing difficulties in finding a health insurance. Continuous growth is expected for the wearables technologies market, reaching 12.6 billion U.S. dollars in 2020 (Business Insider, 2013).
	B	The market power is limited to a small number of large companies who control all servers (cf. Lanier, 2014: “siren servers”).
	C	Elderly people are driving the development, particularly due to health applications (Beauchet et al., 2014). Quantified self-applications turn basically to assisting tools for active aging and elderly healthcare.

Table 5-4 Influence factors and future projections

5.3.7 Morphology Analysis and Scenario Stories

In the following, the future projections are combined in a morphology matrix (see Figure 5-5). For the array of plausible variations, there are projections excluding each other (indicated by red lines), and preferable combinations (given by green lines). For instance, many devices are produced and smart clothes are trending (Projection 3B); these do not meet the low spread of technologies (Projection 2B). And when the hype is over (Projection 3A), no new applications will be designed (Projection 6A). Further, a distributed market, as in Projection 4B, contradicts the few large players as in Projection 6B. As a second step, the projections are combined into three different future scenarios (indicated by numbers 1 to 3 in circles).

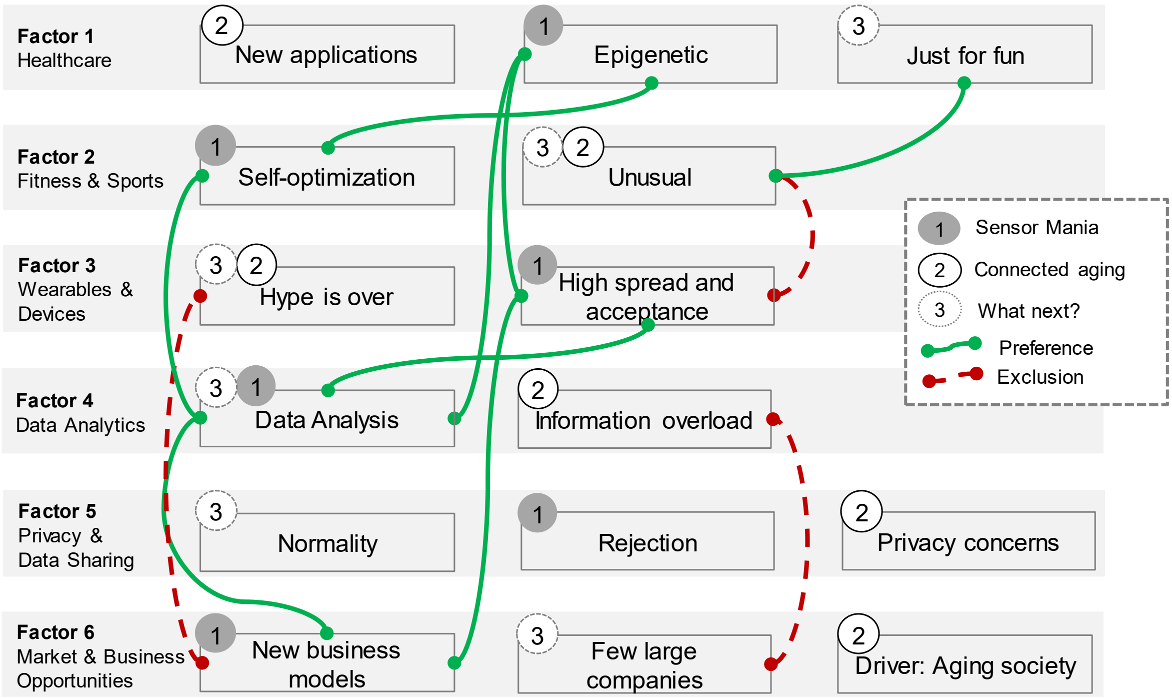


Figure 5-5 Morphology space for *quantified self*

Three short scenario stories are formulated, based on the results of the morphology analysis. As stated previously, a time-frame of five years is aimed at.

Scenario 1: Sensor mania

Quantified self is integrated in our daily life, and everything is connected. Sensors are mass-produced and healthcare applications are far developed, even used in epigenetic applications. New business models are evolving and the data is analyzed for many applications. Of course, each trend has its countertrend, and there is an increasing revolt due to the high sensor-penetration of the daily life. So there are conflicts between manufacturers' business models and users' control of their own personal data.

Scenario 2: Connected aging & health-centric applications

Quantified self is especially used for serious applications, e.g. in the field of health care, but less for sports and fitness. Generally, the hype is over. Technically, the distributed market and the many suppliers make connected applications very difficult. In addition, these developments raise the customers' privacy concerns. Interestingly, one group remained: the elderly. For them, quantified self-applications proved to be very beneficial. So the aging society is a key driver in this scenario, and early detection of diseases is one main concern.

Scenario 3: QS has been established; what next?

A few large enterprises dominate the market and data is analyzed for many purposes. The consumers and users are aware of this fact and have their reservations against this development. The devices are mainly used just for fun, but also in the health care sector. Generally, the hype is over and quantified self is established. The question is: What next?

5.4 Conclusions for Scenario Development

The idea underlying this article is to optimize the process of scenario development. As described in Section 5.2, this especially relates to improving the first step (data collection and aggregation), but also to enhance the second step of developing future scenarios. A crucial issue in scenario processes is to precisely consider the state-of-the-art, and condense it to influence factors for the following process. Web mining, based on Twitter data, was used in this article for a rapid overview of the scenario field and to automate desk research. The results from web mining are summarized through different text mining approaches. A discussion in the scenario team lead to a list of factors facilitating the formulation of future projections. For looking into the future, morphology analysis was applied, and three different scenarios were written. The results of this article are discussed in the following.

5.4.1 Web-based Data Retrieval

This article builds on the idea to use web mining in the context of scenario development to reduce the effort needed for desk research. The basic assumption is that an analysis of Twitter tweets (annotated with the *#quantifiedself*-hashtag) enables the summary and description of the scenario field, giving an impression of what should be considered in the scenarios process.

This form of analysis facilitates insights into the opinions of a larger group than literature review or workshops alone. Instead of relying on input from, e.g., interviews, content can be retrieved from a broad spread of different websites (see, for example, Table 5-2). Of course, the Twitter community is a global combination of people and backgrounds, and their expertise is difficult to validate. But this heterogeneous input is less critical in the context of scenario development because only a starting point is needed of what is currently being discussed. And in comparison, a manual search is strongly influenced by the personal background of the reader, and less information can be processed.

Twitter is only one example for a social media platform applicable to scenario development. Data can be retrieved, depending on the focus or scope of the project; so, using Twitter is only means to an end. For example, other social media platforms might be used in future applications or special services for web scraping. However, one needs to be clear that facing the volumes of today’s information, it is not feasible to capture all information related to a topic, but a part or excerpt. If this excerpt is retrieved faster over channels like Twitter, this is an advantage compared to present approaches. Another point is that not each topic or question can be formulated as a search. The more concrete a topic is, as for example, a specific technology, the easier it is. This means that web mining or other data retrieval cannot be applied in some cases.

5.4.2 Text Mining-based Information Aggregation

Text mining facilitates the structuring and delimitation of the scenario field. Text mining reduces the reading effort by summarizing the content of the websites. More data can thereby be processed and downsides of qualitative approaches can be balanced (e.g., bias in expert selection). Two text mining approaches were applied here: concept mapping and topic modeling. They both have different approaches on how to summarize and aggregate the content (see Kayser and Shala, 2014 for a detailed comparison). Concept Mapping delivers a graphical overview on the scenario field, indicating influence areas and factors. However, *Leximancer,* as a commercial software, remains a black box and it is difficult to intervene in the process. Therefore, a second solution was implemented, topic modeling. In contrast, it is easier to trace and noun phrases can be extracted, but it requires programming skills. In summary, the two solutions complement each other very well, and provide two different perspectives on the text.

Still, future projections need to be searched manually, but the results of the automatic analysis provide some guidance. However, foresight activities should balance the strength of qualitative and quantitative thinking. Text mining delivers a summary and a starting point for the discussions. However, the future projections need to be searched manually, also to resolve potential weaknesses of the data, identify missing aspects and factors, validate the process, and align them with the process objectives.

In contrast to previous work where a set of foresight reports were used, this article used web content and a larger dataset. In future work, this method can be expanded and further developed. For example, the dataset can be analyzed for dynamics in the topic, such as Stelzer et al. (2015) do, or other data analyses can be applied.

5.4.3 Implications for Scenario Development

Desk research and workshops are differentiated for data collection in scenario development (van Notten et al., 2003). The method developed in this article remarkably improves desk research in terms of the resources needed and the overview on the scenario field the scenario team gets. Web mining captures what is actually discussed and text mining aggregates content for scenario development.

For explorative scenarios in particular, the generation phase and receiving input is important (Börjeson et al., 2006). Of course, the results need to be discussed in the scenario team or in workshop formats. Normally, workshops are conducted to shape and influence the direction of the scenarios, especially by formulating the future projections. This article focuses on principally illustrating the method. Owing to the resources in this project, no workshops were conducted in this case, although they are generally possible throughout the process. For example, Franco et al. (2013) distinguish between *procedural* (e.g., collecting and structuring of content) and *discursive* group activities (e.g., debating the strategy). While parts of the procedural activities might be replaced by text mining results, the discursive elements are still mandatory. Workshops gain relevancy especially in the third step, scenario usage. Therefore, this step was skipped in this article. Here, the judgments of the target groups or the customers are necessary to transform the scenarios, for example, into action plans or roadmaps.

The scenario stories in this article are very short because the focus of this article lies on improving desk research, and actually their intention is to illustrate this form of scenario process; the length of the scenario stories can be extended in future work. Morphology analysis is criticized for being useful for a small number of factors and projections but for becoming confusing for more factors (e.g., Kosow and Gassner, 2008). This article contains a small case, but larger cases might be handled with well-designed IT support.

One limitation of quantitative data is that we cannot derive or read the future from it (van der Heijden, 2005; Amer et al., 2013; Pillkahn, 2008). So, data always needs human interpretation. This can be accomplished by qualitative work, for example, by thinking about future projections. Scenario development cannot be automated, especially because stakeholder commitment and judgment are central in foresight exercises (see, e.g., Haegeman et al., 2013). However, a deliberate combination of qualitative and quantitative thinking can save resources. An intention of scenarios is to stretch the mental models and perceptions of individuals (Bradfield, 2008; van der Heijden, 2005). The results from text mining can serve for a first reflection and stretch these mental models by the different perspectives they cover. Therefore, the scenario process as illustrated in this article can serve as a starting point to design new workshop concepts. In addition, these improvements practically make scenario development faster and reduces the number of required workshop days. Eventually, scenarios might be developed within a single day, thanks to the better application of software and IT.

Finally, the opportunities and epistemic value arising from *big data* should be systematically evaluated for foresight and its methods. Therefore, Twitter and web mining were tested for retrieving a starting point for scenario development, and proved to be very helpful. Finally, by this fast-track scenario process, more time is saved for the *real* work: foresight and strategy formulation.

References

Amer, Muhammad; Daim, Tugrul U; Jetter, Antonie. “A review of scenario planning.” *Futures* 46, no. 0 (2013): 23–40.

Anagnostopoulos, Christos; Bason, Tom. “Mapping the First 10 Years with Leximancer: Themes and Concepts in the Sports Management International Journal Choregia.” *CHOREGIA* 11, no. 1 (2015): 25–41.

Angus, Daniel; Rintel, Sean; Wiles, Janet. “Making sense of big text: a visual-first approach for analysing text data using Leximancer and Discursis.” *International Journal of Social Research Methodology* 16, no. 3 (2013): 261–267.

Arthur, Charles. “Wearables: one-third of consumers abandoning devices.” *The Guardian,* April 01, 2014.

Beauchet, Olivier; Launay, Cyrille P; Merjagnan, Christine; Kabeshova, Anastasiia; Annweiler, Cédric. “Quantified self and comprehensive geriatric assessment: older adults are able to evaluate their own health and functional status.” *PloS one* 9, no. 6 (2014): e100636.

Bell, Erica; Seidel, Bastian M. “The evidence-policy divide: a 'critical computational linguistics' approach to the language of 18 health agency CEOs from 9 countries.” *BMC Public Health* 12, no. 1 (2012): 932.

Beuth, Patrick. “Vermessen und verkauft.” *Zeit Online,* April 20, 2015.

Bezold, Clem. “Lessons from using scenarios for strategic foresight.” *Technological Forecasting and Social Change* 77, no. 9 (2010): 1513–1518.

Bird, Steven; Klein, Ewan; Loper, Edward. *Natural language processing with Python*. 1st ed. Beijing, Cambridge [Mass.]: O'Reilly, 2009.

Blei, David M; Lafferty, John D. “A correlated topic model of Science.” *The Annals of Applied Statistics* 1, no. 1 (2007): 17–35.

Blei, David M; Ng, Andrew Y; Jordan, Michael I. “Latent Dirichlet Allocation.” *the Journal of machine Learning research* 3 (2003): 993–1022.

Börjeson, Lena; Höjer, Mattias; Dreborg, Karl-Henrik; Ekvall, Tomas; Finnveden, Göran. “Scenario types and techniques: Towards a user's guide.” *Futures* 38, no. 7 (2006): 723–739.

Bradfield, Ron. “Cognitive Barriers in the Scenario Development Process.” *Advances in Developing Human Resources* 10, no. 2 (2008): 198–215.

Bradfield, Ron; Wright, George; Burt, George; Cairns, George; van der Heijden, Kees. “The origins and evolution of scenario techniques in long range business planning.” *Futures* 37, no. 8 (2005): 795–812.

Bruns, Axel; Burgess, Jean. “Notes towards the Scientific Study of Public Communication on Twitter.” *Science and the internet (*2012): 159–169.

Business Insider. “Wearable device market value from 2010 to 2018 (in million U.S. dollars).” 2013. http://www.statista.com/statistics/259372/wearable-device-market-value/.

Cameron, David; Finlayson, Amalie; Wotzko, Rebecca. “Visualising Social Computing Output: Mapping Student Blogs and Tweets.” In *Social Media Tools and Platforms in Learning Environments,* edited by Bebo White, Irwin King and Philip Tsang. Berlin, Heidelberg: Springer, 2011: 337–350.

Cretchley, Julia; Gallois, Cindy; Chenery, Helen; Smith, Andrew E. “Conversations between carers and people with schizophrenia: a qualitative analysis using leximancer.” *Qualitative health research* 20, no. 12 (2010): 1611–1628.

Davies, Islay; Green, Peter; Rosemann, Michael; Indulska, Marta; Gallo, Stan. “How do practitioners use conceptual modeling in practice?” *Data & Knowledge Engineering* 58, no. 3 (2006): 358–380.

Feldman, Ronen; Sanger, James. *The text mining handbook: Advanced approaches in analyzing unstructured data*. Cambridge: Cambridge University Press, 2008.

Franco, L. A; Meadows, Maureen; Armstrong, Steven J. “Exploring individual differences in scenario planning workshops: A cognitive style framework.” *Technological Forecasting and Social Change* 80, no. 4 (2013): 723–734.

“Gartner Predicts By 2017, 30 Percent of Smart Wearables Will Be Inconspicuous to the Eye.” Gartner press release. Stamford, December 10, 2014. http://www.gartner.com/newsroom/id/2941317.

Gausemeier, Jürgen; Fink, Alexander; Schlake, Oliver. *Szenario-Management: Planen und Führen mit Szenarien*. 2nd ed. München, Wien: Hanser, 1996.

Gibbs, Samuel. “Forget smartwatches - smartclothes are the future, analysts say.” *The Guardian,* November 18, 2014.

GIbbs, W. W. “Medicine gets up close and personal.” *NATURE* 506, no. 7487 (2014): 144–145.

Glenn, Gerome C; The Futures Group International. “Scenarios.” In *Futures research methodology*. 3rd ed., edited by Jerome Clayton Glenn and Theodore J. Gordon. Washington, DC: The Millennium Project, 2009: 1–25.

Godet, Michel. “Scenarios and strategies. A toolbox for problem solving.” Cahiers du LIPS, 1997.

Grunwald, Armin. *Technikfolgenabschätzung: Eine Einführung*. 2nd ed. Berlin: Ed. Sigma, 2010.

Haegeman, Karel; Marinelli, Elisabetta; Scapolo, Fabiana; Ricci, Andrea; Sokolov, Alexander. “Quantitative and qualitative approaches.” *Technological Forecasting and Social Change* 80, no. 3 (2013): 386–397.

van der Heijden, Kees. *Scenarios: The Art of Strategic Conversation*. Chichester, England: John Wiley & Sons, 2005.

Hirsch, Sven; Burggraf, Paul; Daheim, Cornelia; Luis Cordeiro, José. “Scenario planning with integrated quantification – managing uncertainty in corporate strategy building.” *foresight* 15, no. 5 (2013): 363–374.

Java, Akshay; Song, Xiaodan; Finin, Tim; Tseng, Belle. “Why we twitter: understanding microblogging usage and communities.” Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis: ACM, 2007: 56–65.

Kayser, Victoria; Shala, Erduana. “Generating Futures from Text: Scenario Development using Text Mining.” 5th International Conference on Future-Oriented Technology Analysis (FTA) - Engage today to shape tomorrow. Brussels, Belgium, 2014.

Kosow, Hannah; Gassner, Robert. *Methoden der Zukunfts- und Szenarioanalyse: Überblick, Bewertung und Auswahlkriterien*. Berlin: IZT, 2008.

Kuosa, Tuomo. *The Evolution of Strategic Foresight: Navigating Public Policy Making*. Farnham: Ashgate Publishing Ltd, 2012.

Kwak, Haewoon; Lee, Changhyun; Park, Hosung; Moon, Sue. “What is Twitter, a social network or a news media?”. Proceedings of the 19th International Conference on World Wide Web. Raleigh, NC, USA: ACM Digital Library, 2010: 591–600.

Lanier, Jaron. *Wem gehört die Zukunft?: Du bist nicht der Kunde der Internetkonzerne, du bist ihr Produkt*. 2nd ed. Hamburg: Hoffmann und Campe, 2014.

Leximancer. “Leximancer Manual: Version 4.” 2011.

Liesch, Peter W; Håkanson, Lars; McGaughey, Sara L; Middleton, Stuart; Cretchley, Julia. “The evolution of the international business field: a scientometric investigation of articles published in its premier journal.” *Scientometrics* 88, no. 1 (2011): 17–42.

Liu, Bing. *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data*. 2nd ed. Berlin, Heidelberg: Springer-Verlag Berlin Heidelberg, 2011.

Mietzner, Dana; Reger, Guido. “Advantages and disadvantages of scenario approaches for strategic foresight.” *International Journal of Technology Intelligence and Planning* 1, no. 2 (2005): 220–239.

Miner, Gary. *Practical text mining and statistical analysis for non-structured text data applications*. 1st ed. Waltham, MA: Academic Press, 2012.

van Notten, Philip W. F; Rotmans, Jan; van Asselt, Marjolein B. A; Rothman, Dale S. “An updated scenario typology.” *Futures* 35, no. 5 (2003): 423–443.

O'Brien, Frances A; Meadows, Maureen. “Scenario orientation and use to support strategy development.” *Scenario Method: Current developments in theory and practice* 80, no. 4 (2013): 643–656.

Olson, Parmy. “Fitbit Data Now Being Used In The Courtroom.” *Forbes,* November 16, 2014.

Pillkahn, Ulf. *Using Trends and Scenarios as Tools for Strategy Development*. Hoboken: John Wiley & Sons; Wiley-VCH, 2008.

Poser, Claudia; Guenther, Edeltraud; Orlitzky, Marc. “Shades of green: Using computer-aided qualitative data analysis to explore different aspects of corporate environmental performance.” *Journal of Management Control* 22, no. 4 (2012): 413–450.

PWC. “The Wearable Future: Consumer Intelligence Series.” 2014.

Raford, Noah. “Crowd-sourced Collective Intelligence Platforms for Participatory Scenarios and Foresight.” *Journal of Futures Studies* 17, no. 1 (2012): 112–128.

Raford, Noah. “Online foresight platforms: Evidence for their impact on scenario planning & strategic foresight.” *Technological Forecasting and Social Change* 97 (2015): 65–76.

Řehůřek, Radim; Sojka, Petr. “Software Framework for Topic Modelling with Large Corpora.” Proceedings of LREC 2010 Workshop 2010. Valletta, Malta, 2010: 46–50.

Reibnitz, Ute. *Szenario-Technik: Instrumente für die unternehmerische und persönliche Erfolgsplanung*. Wiesbaden: Gabler, 1991.

Rooney, D; McKenna, B; Barker, J. R. “History of Ideas in Management Communication Quarterly.” *Management Communication Quarterly* 25, no. 4 (2011): 583–611.

Salton, Gerard. *Automatic Text Processing: The transformation, analysis, and retrieval of information by computer*. Reading, Mass.: Addison-Wesley, 1988.

Siegel Bernhard, Tara. “Giving Out Private Data for Discount in Insurance.” *The New York Times,* April 08, 2015.

Smith, Andrew E; Humphreys, Michael S. “Evaluation of unsupervised semantic mapping of natural language with Leximancer concept mapping.” *Behavior Research Methods* 38, no. 2 (2006): 262–279.

Stelzer, Birgit; Meyer-Brötz, Fabian; Schiebel, Edgar; Brecht, Leo. “Combining the scenario technique with bibliometrics for technology foresight:: The case of personalized medicine.” *Technological Forecasting and Social Change* 98 (2015): 137–156.

Stockwell, Paul; Colomb, Robert M; Smith, Andrew E; Wiles, Janet. “Use of an automatic content analysis tool: A technique for seeing both local and global scope.” *International Journal of Human-Computer Studies* 67, no. 5 (2009): 424–436.

Swan, Melanie. “Sensor Mania!: The Internet of Things, Wearable Computing, Objective Metrics, and the Quantified Self 2.0.” *Journal of Sensor and Actuator Networks* 1, no. 3 (2012): 217–253.

Tozer, Daniel. “Legal: The laws and regulations of wearable devices.” *WearableTech,* September 10, 2015.

Vidal, Leticia; Ares, Gastón; Machín, Leandro; Jaeger, Sara R. “Using Twitter data for food-related consumer research: A case study on “what people say when tweeting about different eating situations”.” *Food Quality and Preference* 45 (2015): 58–69.

Weintraub, Karen. “Quantified self: the tech-based route to a better life?” *BBC Future,* January 03, 2013.

Yarowsky, David. “Unsupervised word sense disambiguation rivaling supervised methods.” Proceedings of the 33rd annual meeting on Association for Computational Linguistics. Cambridge, Massachusetts, USA, 1995: 189–196.

Yau, Chyi-Kwei; Porter, Alan; Newman, Nils; Suominen, Arho. “Clustering scientific documents with topic modeling.” *Scientometrics* 100, no. 3 (2014): 767–786.

Zwicky, Fritz. “Morphological astronomy.” *The observatory* 68 (1948): 121–143

.

III. Final Conclusion

This thesis examines the contribution of text mining for foresight and its methodology. Different data sources, processes and text mining approaches are implemented in existing foresight methods. The starting point for this thesis was the observation that recent foresight methods are often based on literature analysis, patent and publication data or expert opinions, but make few use of other data sources. In times of big data, many other options exist, in particular using web content. Much textual data - such as news, social media*,* or websites - is not considered systematically in foresight activities or analyzed in an automatic or comprehensive manner. In addition, most foresight methods do not make use of text mining. Therefore, this thesis shows several options how to apply text mining in the context of foresight.

Results of the five Articles

This thesis argues the contribution of text mining on a methodological respectively operational level concerning its use in the context of foresight. The first article sets a framework for the following four articles by introducing foresight and text mining and summarizing the state-of-the-art of recent work combining both. Here, the relevancy of text mining for foresight is discussed and showed to be very promising.

The second article introduces an adapted roadmapping process that integrates text mining illustrated for scientific publication data. This approach divides roadmapping in four steps and proposes adequate text mining methods within each. Continuous feedback loops between the strategic work of roadmapping and the analysis with text mining integrate firm-external developments and align them with internal deliberations. The underlying idea is that results of text mining reflect, check or validate intermediate results from the ongoing foresight activity. Furthermore, this framework is adaptable to other data sources such as social media, news reporting or patents. The enhanced early recognition of change by text mining and the parallel strategic adaptation can improve the dynamic capabilities of firms.

The third article and develops a framework for the automatic comparison of textual data exemplified on news articles and abstracts of scientific publications. This article uses the innovation system as a conceptual base; so, text mining assists in better understanding ongoing changes and developments and estimates their systemic implications. For the comparison of public and scientific discourses, parallels can be drawn to technology lifecycles and spread of different technologies. News articles have been analyzed with (qualitative) content analysis for long, but text mining can process larger volumes of text in less time. Principally, this method enables to examine questions related to the spread of the topic, innovation diffusion, or technology acceptance. However, the results deliver hypothesis that need to be proven by other methods.

The fourth article shows the use of data not commonly considered in foresight exercises on the example of Twitter data. As the literature overview has shown, social media and web data are rarely used in foresight and particularly not as a data source. However, when *new* data is used in foresight, first its strength and limitations should be elaborated. As the results of this article show, platforms as Twitter enable the involvement of stakeholders not considered in foresight otherwise (e.g., specific interest groups, young people) and especially a larger number of views and opinions can be processed. The variety and breadth of the gathered content is not possible with classical methods such as interviews or workshops. Twitter data indicates if there is a debate about a certain issue, what is discussed and how. Owing to limitations of Twitter data as being very diverse and not representative, the results should be weighted up with other data and integrated in a larger foresight framework. As the considerations in this article show, many options evolve for conducting research with Twitter in future work such as web mining or examining technology spread and acceptance.

Building on the results of the fourth article, the fifth article introduces web-based scenario development. Thereby, the effort for desk research is reduced by using Twitter data as starting point for conducting web mining based on the links contained in the tweets. The results of web mining are aggregated with text mining and support different steps of the scenario process. By this methodological extension, a multitude of opinions is considered and the time effort for summarizing the scenario field is reduced.

Using Text Mining for Foresight: Final Assessment

The following summarizes the main contributions of text mining for foresight but especially discusses some of its key limitations. To begin with, text mining helps in exploiting the steadily rising volume of (textual) data. Among the main contributions are the larger numbers of opinions and statements that can be integrated in foresight using text mining. For instance, by using Twitter the views of more people are analyzed than by conducting a small number of workshops. The same holds for the number of news articles that can be processed with text mining in contrast to content analysis. Further on, text mining analyzes data that cannot be quantitatively processed in foresight otherwise (e.g., Twitter, news articles, web mining), especially not in this volume and size. To anticipate current developments and integrate an objective external view, text mining is most promising.

However, text mining has limitations as well. Text mining analyses textual data, especially based on word frequencies and word relations. This ignores images and figures – and 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). For example, text mining is not sufficient for sentiment analysis on news articles due to the characteristics of the written text, or the alignment of the results of text mining with functions of innovation systems (see Section 3).

As already indicated in the first article, domain knowledge is inevitable for conducting data analysis to understand and interpret the results. Furthermore, text mining cannot replace reading and algorithms handling data lead to a loss of meaning and context. This explains why some research questions still require qualitative and a more in-depth analysis than statistical approaches as text mining can deliver. As for example the fifth article on web mining shows, a rapid overview on the field as a summary is delivered, but a more detailed level (e.g., statistics, facts, numbers) is manual work. The results of text mining are, in the end, an approximation.

As a matter of fact, text mining cannot automate foresight due to the collaborative character of foresight but offers a reflection and comparison to internal views. In any case, foresight needs stakeholder interaction and involvement (see e.g., Haegeman et al., 2013) - an aspect that cannot be automated because at least setting the process objectives and interpreting the results are required (see, e.g., de Miranda Santo et al., 2006).

Furthermore, there are technical restrictions for the application of text mining. Text mining solutions must be designed and implemented. As stated in the first article, an own adaptable framework meets the requirements best (e.g., adaptable to further data sources, modular design, extendable). This adaptable framework requires IT capabilities, in most cases financial resources, and brings a high learning effort. Each time, text mining brings a manual effort for the initial set up. To process a further type of textual data, adaptions of the interfaces are necessary because the structure of each type of dataset varies. However, commercial software is an option as well.

Implications for Foresight

Designing foresight applications and methods, the added value of text mining is that data sources can be accessed not used so far and more data can be processed and better analyzed. Foresight has a broad spectrum ranging from micro to macro perspective and, as this thesis shows, questions at different scopes of foresight can be addressed with text mining. This reaches from examining systemic links and the function of innovation systems to enhancing the dynamic capabilities of firms. As shown by this thesis, different combinations of data sources, text mining approaches, foresight process requirements (e.g., scope, time frame, topic) and foresight methods are most promising and lead to a wide range of options for future work. These four building blocks can be interchangeably combined and extend the spectrum of foresight methods (see Figure III-1). The modular character of foresight now encompasses further components and many new applications on corporate, organizational, societal or political level can be built in future work.

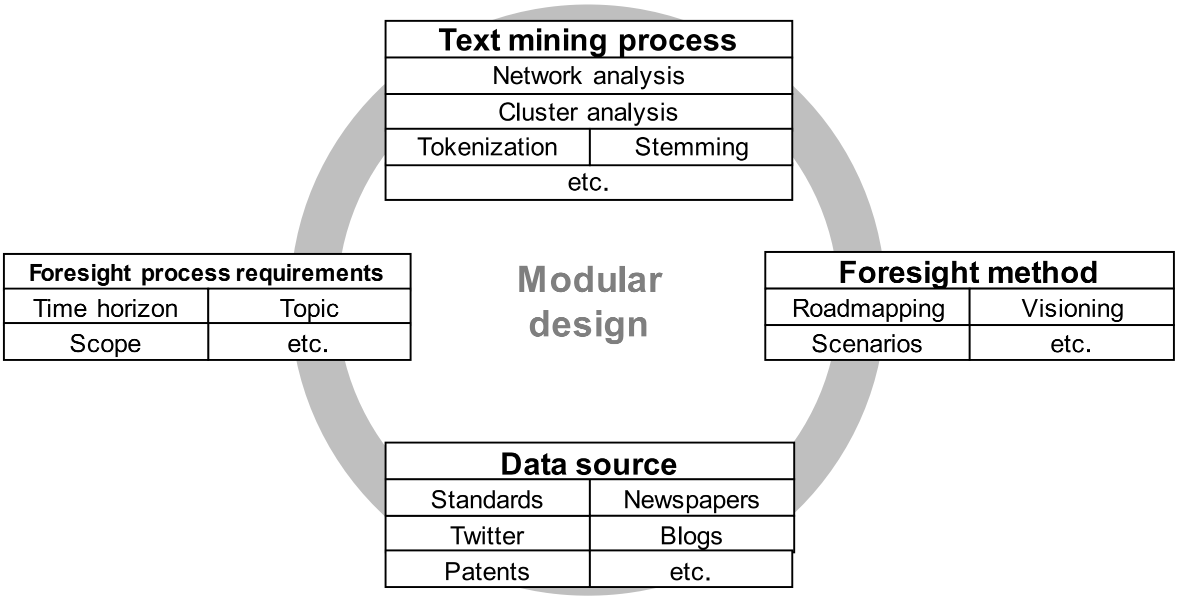


Figure III-1 Modular foresight applications using text mining

Processing more and *new* data enables to integrate more views and stakeholder positions, and, thereby, extends the knowledge base of foresight. However, a principal 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 only from data, because the farer we reach into the future, the less accurate this data gets (Amer et al., 2013; van der Heijden, 2005). In any case, interpretation and alignment with the specific requirements is mandatory 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 that depend on the assumptions made today. Anyway, foresight is less about precision and accuracy but thinking about different futures and the options we have today. This is accomplished by qualitative methods for the long-term view that additionally strengthen the collaborative character of foresight. In the end, this means the farer we reach into the future, the more qualitative foresight gets as highlighted in Figure III-2. Finally, qualitative and quantitative approaches complement each other and, therefore, should be combined (see, e.g., Amer et al., 2013).

These considerations are especially valid for explorative foresight approaches which proceed from now into the future. In this case, it is essential to have a profound understanding of present developments based on data analysis when thinking about complex futures. A broad and comprehensive data basis summarizing the state-of-the-art is the best possible starting point for further steps in the foresight process. By summarizing the "*where are we know*" with text mining, more plausible future paths can be drawn. Thinking about the future based on this analysis provides a solid base for decisions made today and more robust strategies can be derived. Furthermore, in normative settings as, for example, roadmapping, future strategies are developed according to future visions and objectives. In these cases, explorative and normative thinking can be combined and text mining contributes an explorative view in a structured manner.

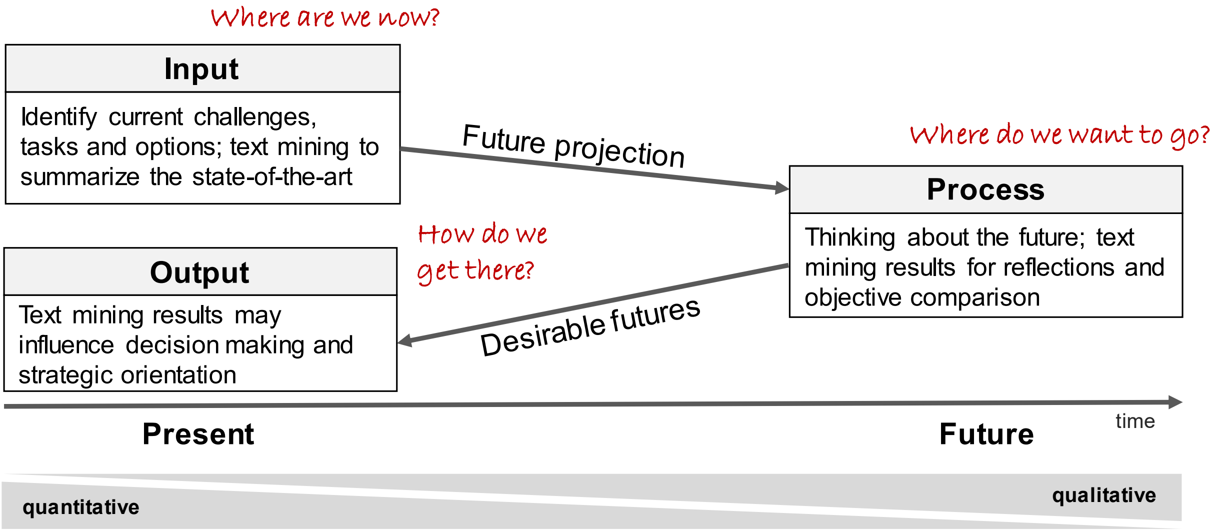

Where are we now?
Input
Identify current challenges, tasks and options; text mining to summarize the state-of-the-art
Future projection
Where do we want to go?
Process
Thinking about the future; text mining results for reflections and objective comparison
How do we get there?
Output
Text mining results may influence decision making and strategic orientation
Desirable futures
Present
Future
time
quantitative
qualitative

Figure III-2 Explorative foresight using text mining: balance of qualitative and quantitative approach

For the process of foresight, this means that the results from text mining serve for reflecting existing ideas and generating an awareness for trends and developments that maybe were not considered before. Thereby, text mining contributes at different points in the foresight process (see also Section 1).

First, exploring and identifying relevant aspects in an objective manner is eased by text mining. This objective and structured summary of the state-of-the-art solves problems of other forms of getting input, such as workshops. The latter might be dominated from single opinions or people that want to push the discussion in 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 for summarizing the considered field. So, more content and thereby more opinions and views can be processed and considered.

Second, for exploring the future, text mining highlights recent trends, contributes an external perspective and serves for reflections. This serves as a starting point for discussing possible futures promoting a creative discourse, in particular by hinting towards former disregarded aspects. In addition, results of text mining may reflect or validate intermediate results from the ongoing foresight activity and, thereby, contribute to the generation of future knowledge.

Finally, results from text mining are valuable to quantify and underline statements made or quantify strategic choices. This supports decision making and in achieving future objectives.

Limitations of this Thesis and Future Research Directions

This thesis, in particular, provides a conceptual outline of how to improve foresight by using text mining. The articles underlying this thesis are proofs of concept and illustrate methodological extensions. In order to assess further the added value of this thesis, more tests and evaluations are necessary. Nevertheless, the conceptual effort for designing and implementing text mining-based foresight methods is not to be underestimated. Examining the relevance of text mining for foresight is the key aim and main concern of this thesis. A systematic evaluation was, from the beginning, beyond the scope of this thesis. Here, the focus is on the integrative step on the methodological level illustrated on smaller datasets. So far, methods are outlined and applications based on different textual data are implemented.

Finally, it can be concluded that text mining for foresight is a rewarding combination and should be pursued in future. However, evaluating the methods as proposed in this thesis, 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. More personal and financial resources will enable larger applications than those realized in this thesis. However, some of the strengths and weaknesses of the methods developed in this thesis can be assessed better.

Concerning the future of foresight and text mining, there is much space for their co-evolution. During the last few years, foresight could increase its acceptance and spread. Especially the volume of data will increase in the future. So, the modular concept, as illustrated in Figure III-1, can be developed further and be used as a starting point for many new foresight applications. In future works, foresight methods, textual data sources and text mining approaches can be combined to different foresight processes in order to design new applications such as real-time foresight or web-based roadmapping.

In any case, solutions to deal with increasing volumes of data, such as text mining, are necessary and very relevant in our present time to exploit these information sources and gain relevant insights. *Big data* offers potentials as a 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 data sources. However, as stated before, the strength and weaknesses of this data, in combination with text mining, should be clear and evaluated further in future work.

A further point is gaining more experience in processing web data. Techniques such as web mining enable timely 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 implies many new options, especially in 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 as social media or user-generated content are rarely considered in foresight to capture societal points of view. Using text mining, a larger set of views and opinions can be analyzed for a rapid overview on issues discussed as shown in this thesis on the example of Twitter of news articles. This enables new forms of stakeholder integration and engagement in foresight processes. Therefore, more research effort should be spent here.

This thesis illustrates different ways of accessing and aggregating today's volumes of data and information 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 weight up 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 deployed in foresight.

References

Amer, Muhammad; Daim, Tugrul U; Jetter, Antonie. “A review of scenario planning.” *Futures* 46, no. 0 (2013): 23–40.

Boyd, Danah; Crawford, Kate. “Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon.” *Information, communication & society* 15, no. 5 (2012): 662–679.

Grimmer, Justin; Stewart, Brandon M. “Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts.” *Political Analysis (*2013): 1–31.

Haegeman, Karel; Marinelli, Elisabetta; Scapolo, Fabiana; Ricci, Andrea; Sokolov, Alexander. “Quantitative and qualitative approaches.” *Technological Forecasting and Social Change* 80, no. 3 (2013): 386–397.

van der Heijden, Kees. *Scenarios: The Art of Strategic Conversation*. Chichester, England: John Wiley & Sons, 2005.

de Miranda Santo, Marcio; Coelho, Gilda M; dos Santos, Dalci M; Filho, Lélio F. “Text mining as a valuable tool in foresight exercises: A study on nanotechnology.” *Technological Forecasting and Social Change* 73, no. 8 (2006): 1013–1027.

Pillkahn, Ulf. *Using Trends and Scenarios as Tools for Strategy Development*. Hoboken: John Wiley & Sons; Wiley-VCH, 2008.