Technological Forecasting & Social Change 115 (2017) 348–357

Contents lists available at ScienceDirect

Technological Forecasting & Social Change

Comparing public and scientific discourse in the context of innovation systems

Victoria Kayser

Fraunhofer Institute for Systems and Innovation Research, Breslauer Strasse 48, 76131 Karlsruhe, Germany
Technische Universität Berlin, Chair of Innovation Economics, Marchstraße 23, 10623 Berlin, Germany

ARTICLE INFO

Article history:
Received 30 September 2015
Received in revised form 11 July 2016
Accepted 2 August 2016
Available online 25 August 2016

Keywords:
Innovation system
Text mining
Foresight
Media analysis
Publications analysis
Future technology analysis

ABSTRACT

Innovation as a systemic process is not only driven by science and technology but has diverse sources. While there are (numeric) indicators to map S&T developments such as patents, publications or standards, new indicators are required to map other areas of the innovation system. In this regard, one option is the examination of news reporting. News is a recognized channel for innovation diffusion and plays an important role in informing society. To contrast changes and developments in science and society, specifically the link between both is addressed in this article by comparing the content of news articles and scientific publications. Thus, the aim of this article is to first argue the benefit of integrating the media in the innovation system debate because of its recognized role in innovation diffusion and to develop a methodology to automatically compare scientific and media discourses. To process the volume of textual data according to a common analytical scheme, a text mining framework has been developed. The results offer valuable input for examining the present state of themes and technologies and, thereby, support future planning activities.

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. For the long-term observation of thematic and technical developments, an analysis should not only address one area of the innovation system (i.e., science) but should consider further parts (i.e., society).

Innovation as a systemic process is not only driven by science and technology but has diverse sources. While there are (numeric) indicators to map S&T developments such as patents, publications or standards, new indicators are required to map other areas of the innovation system. In this regard, one option is the examination of news reporting to map societal discourse. News is a recognized channel for innovation diffusion and plays an important role in informing society. 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.

To contrast changes and developments in science and society, specifically the link between the two is addressed in this article by comparing the content of news articles and scientific publications. Publication data is a commonly used source for examining scientific progress (e.g., Leydesdorff and Milojević, 2015). This work will explore if it is possible to (automatically) recognize changes and focus in reporting of both, science and news. This potentially enables inferences on the state of a technology or its societal acceptance for the comparison of subsystems. To process the volume of textual data according to a common analytical scheme, a text mining framework has been developed. Currently, there is no methodology for the (automatic) comparison of news articles and scientific publications but theoretical discussions (e.g., Franzen et al., 2012).

Thus, the aim of this article is to first argue the benefit of integrating the media in the innovation system debate because of its recognized role in innovation diffusion and introducing an adapted innovation system model as conceptual framework. Then a method is developed to automatically compare scientific and media discourses based on *textual data*. It is examined if differences in the discourse can be measured and mapped based on news articles and scientific publication abstracts. Therefore, a framework based on text mining is developed. This might illustrate 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

E-mail address: mail@vkayser.de.

http://dx.doi.org/10.1016/j.techfore.2016.08.005
0040-1625/© 2016 Elsevier Inc. All rights reserved.

(Section 2). Then, Section 3 describes the framework of analysis while Section 4 introduces the three case studies. In Section 5, the results are discussed and final conclusions are drawn.

2. Foundations

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

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 defined 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.

2.2. Integrating media in the innovation system debate

While it is commonly agreed that innovation needs to be viewed systemic, the society as a framework or media as a distribution channel are no explicit elements of prominent definitions of innovation systems (Waldherr, 2008, 2012). For this reason, this article discusses the role of 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). Thereby mass media has certain functions in society (e.g., 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 and map the dynamics and processes of diffusion introduces an adapted model.

Waldherr (2012) argues that mass media is an important intermediary in the triangle of politics, economy, and research (see Fig. 2-1). 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.

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 and the interrelation between societal and scientific discourse is analyzed in more detail in this work. So this article examines if it is possible to automatically compare news articles and publication abstracts and develops a method for this purpose. These two data sources are of comparable language and text quality and will therefore be used for a principal investigation. Of course, if the results show to be useful, the method can be developed further and further data can be integrated in future work.

3. Methodology: comparing datasets

Focus of the following section is to introduce a method that is able to automatically compare scientific and public discourse. A manual approach is too time consuming due to the size of the data sets. 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 work, after which the analysis framework is introduced.

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 work 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)

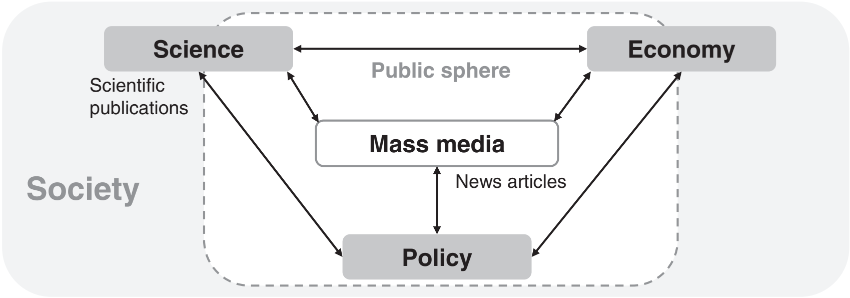

Fig. 2-1. Adaptation of the innovation system (own illustration with reference to Waldherr, 2012).

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).

3.2. 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 because of its combination of different news sources and its searchable database.

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 summarizes 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. 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.

As described in Fig. 3-1, first, the number of records per year in the two datasets, news articles and scientific publications, has been compared (time series). 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 text fields are analyzed in more detail using noun phrase extraction (as described in Section 3.3.1). This is motivated by the question of which aspects the datasets focus on and the coverage and volume of reporting. Summaries of each text source are visualized as term networks Section 3.3.2 and he extracted nouns are matched and visualized as pie bubble chart (see Section 3.3.3).

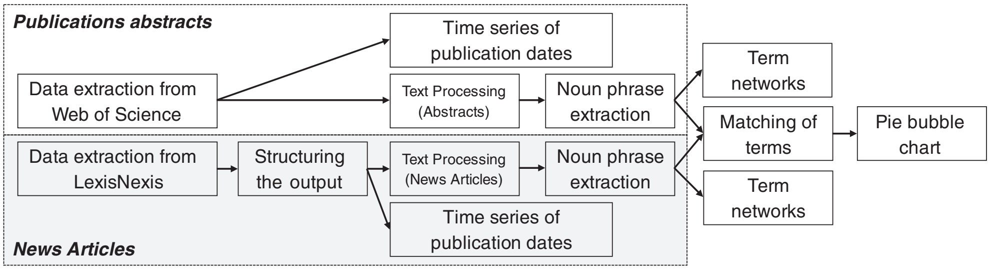
Publications abstracts
Data extraction from Web of Science
Time series of publication dates
Text Processing (Abstracts)
Noun phrase extraction
Term networks
Matching of terms
Pie bubble chart
Term networks
Data extraction from LexisNexis
Structuring the output
Text Processing (News Articles)
Noun phrase extraction
Time series of publication dates
News Articles

Fig. 3-1. Process model for comparing data sets.

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.

3.3.1. Processing textual data

To summarize covered content, 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. Then, *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.

Another option for summarizing the textual data sets are techniques for text summarization (e.g., Das and Martins, 2007; Batista et al., 2015). However, the short summaries from multi-document summarization (only few sentences) are too short for the research objective addressed in this article and not as easily comparable as extracted terms. Moreover, single-document summarization delivers too many summaries to contribute to this comparative approach. In addition, the summarizes are not comparable in an automatic or numeric manner. So term extraction and different visualization techniques are applied in the following.

3.3.2. Term networks

For a first overview on the contained terms and their interrelations, term networks for the 100 most frequent terms are drawn for each dataset as an initial overview of the terms and their co-occurrence. This analysis builds on the terms extracted in the previous step. The networks illustrate how the terms are interconnected and, therefore, depend on each other. In contrast to, e.g., 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).

3.3.3. Matching and comparing the datasets

Next, the terms contained in publication abstracts and news articles are matched to identify common and unique terms. This illustrates where aspects are addressed more and which differences exist. Technically, this is realized in SQL by comparing term frequencies and occurrence. The results are mapped as *pie bubble charts* for a better overview. 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-

Table 4-1
The three datasets.

		Scientific publications	News articles
		Web of Science	*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

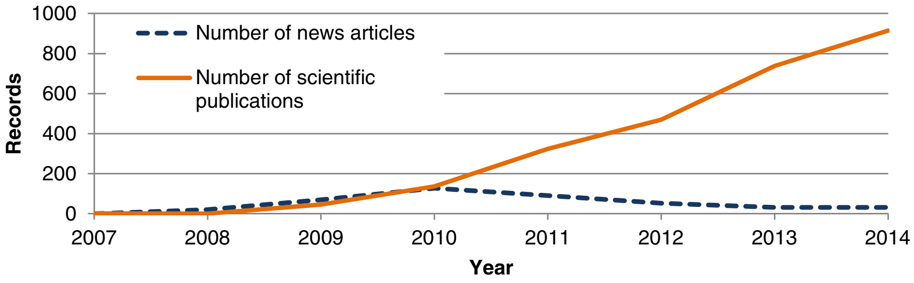

Fig. 4-1. Cloud computing: Number of records.

axis (avoiding overlaps of bubbles), the x-axis represents the degree of inclusion in the news (left side) or in scientific publications (right side). This 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* enable 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*.

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 4-1 describes the data and gives an overview on the searches and their results.

4.1. Cloud computing

Cloud computing (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. Fig. 4-1 gives an overview of 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.

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 Fig. 4-2 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

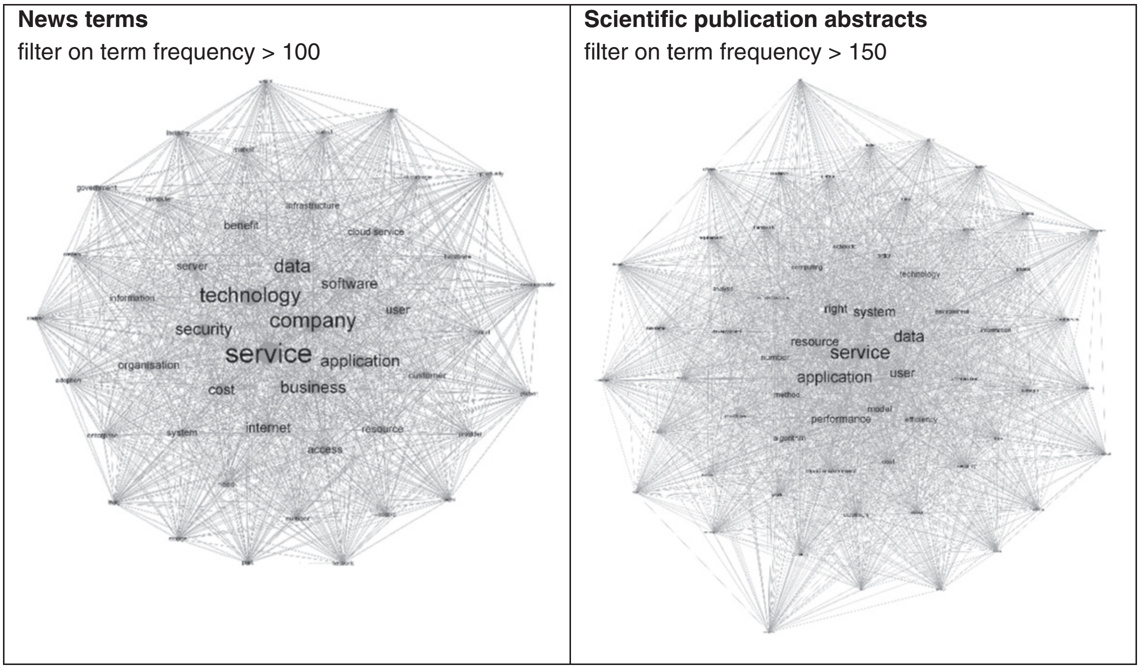

Fig. 4-2. Cloud computing: Network of terms.

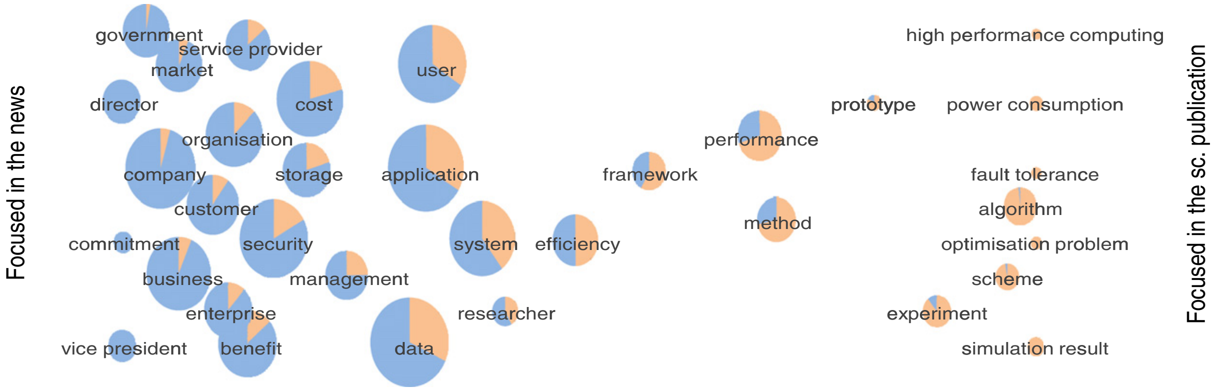

Fig. 4-3. Cloud computing: Pie bubble chart (selection of terms).

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*).

Next, the *pie bubble chart* directly compares the frequency of terms in the two datasets (Fig. 4-3). 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 articles 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.

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 Fig. 4-3 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*.

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 Fig. 4-4 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).

Next, the texts are processed. The 100 most frequent terms per dataset are visualized in term networks (Fig. 4-5). 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*).

The *pie bubble chart* as illustrated in Fig. 4-6 shows that terms occurring only in the abstracts are rare (e.g., *phenyl, fluorescence spectra*), with the exception of *electron transfer* and *water oxidation.* The abstracts are dominated by (scientific) terms such as *absorption and oxidation*. Terms such as *professor* and *university* occur only in the news. This indicates that the news reports a lot about scientific work and progress, especially in the context of energy generation. This especially highlights, that the news report on a higher abstraction level as e.g., energy generation in general and observe what is going on in science. An own (societal) discourse on the topic is not noticeable from this form of analysis, opposed to the next case on vegan diet.

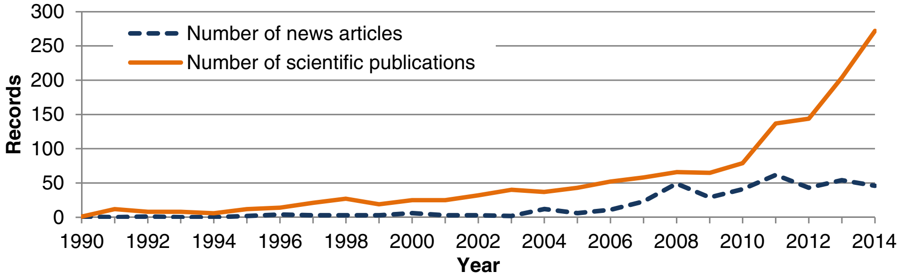

Fig. 4-4. Artificial photosynthesis: Number of records.

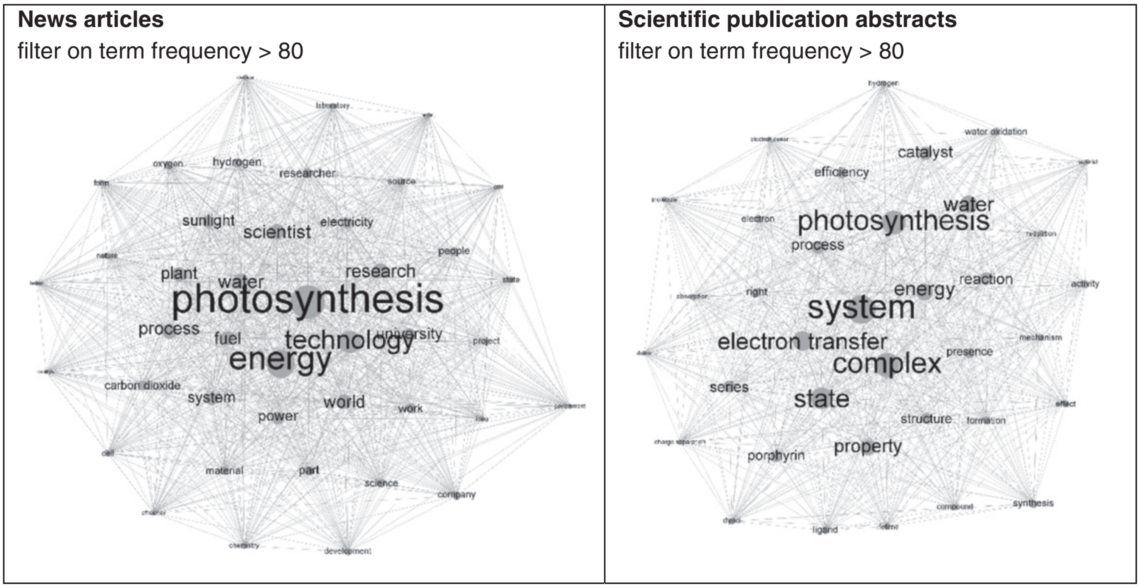

Fig. 4-5. Artificial photosynthesis: Network of terms.

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 (Fig. 4-7). 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.

The texts (news articles and abstracts) are processed and for an overview on the thematic focus in each dataset, term networks are drawn (Fig. 4-8). 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*).

As Fig. 4-9 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. This comparative analysis shows that there are mostly two different issues discussed and there is few overlap.

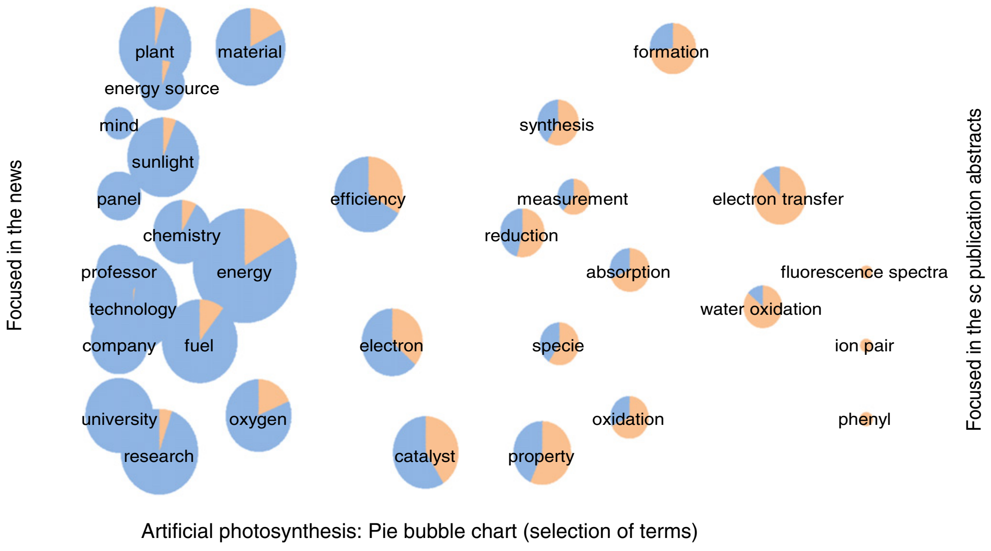

Fig. 4-6. Artificial photosynthesis: Pie bubble chart (selection of terms).

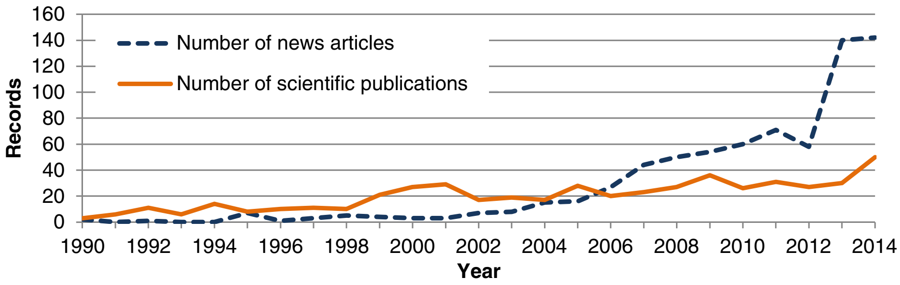

Fig. 4-7. Vegan diet: Number of records.

5. Discussion and conclusion

This article examines if it is possible to automatically compare news articles and publication abstracts and develops a method for this purpose. Furthermore, 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 for 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 the context of foresight. 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 extension of the numeric overview on the number of articles respectively publications illustrates how the terms are connected and gives a rapid 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. Thereby, the exchange can, to a certain degree, be mapped and 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 2010. On the other hand, in the case of *vegan diet,* the media reports a lot, but about different things than science. However, as Hekkert and Negro (2009) conclude, many of these knowledge diffusion processes are not explicitly noticeable and therefore cannot be mapped and recognized. The results of this method allow certain conclusions, but there remains a great deal on the level of hypothesis that should be proven by additional examinations. However, 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). 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

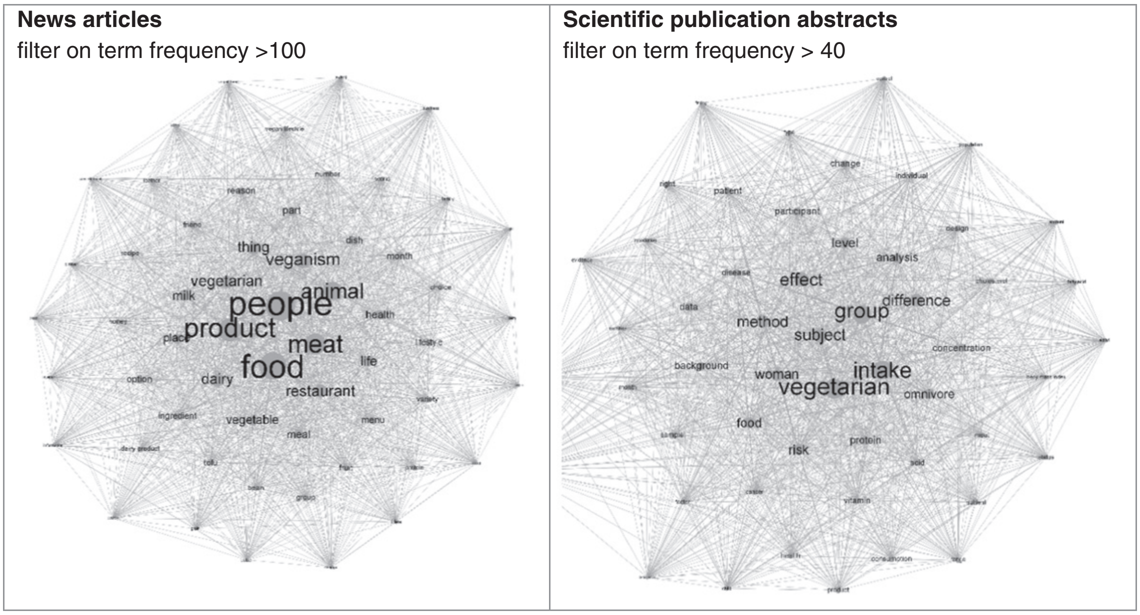

Fig. 4-8. Vegan diet: Network of terms.

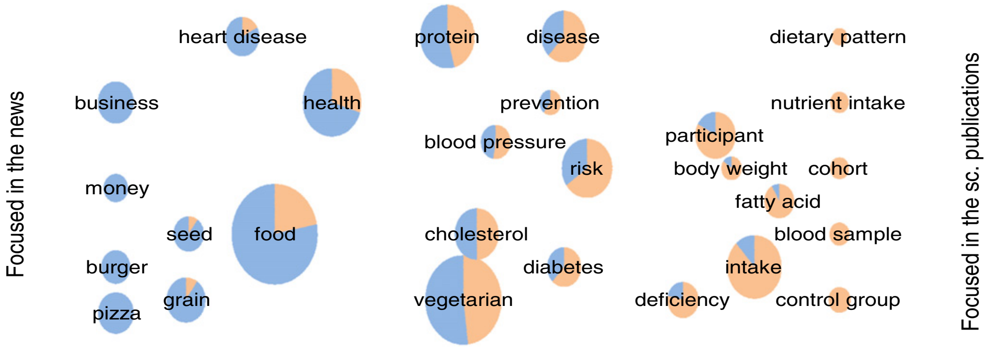

Fig. 4-9. Vegan diet: Pie bubble chart (selection of terms).

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 lays a basis that can be developed in various directions in future. This especially relates to five 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. Second, 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 as illustrated in Fig. 5-1. The analysis of social media and user-generated content is an option to further map society, e.g. based on the analysis of Twitter data. This is, of course, an excellent point for future research to build on the framework respectively method developed here.

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, on the one hand, 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. On the other hand, the link to the maturity of a technology might be further explored by specifically examining when in the innovation lifecycle innovations are discussed in the media. Focus of this article was the principal examination of a framework for the automatic comparison of texts and the technical realization. Fourth, additional (qualitative) methods might be used to validate the hypothesis and observations. And, fifth, future work could stronger link the framework as described in this article to communication science and media theory what was not the focus here.

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, media might be recognized as an element in innovation systems due to its 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 2.2 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,

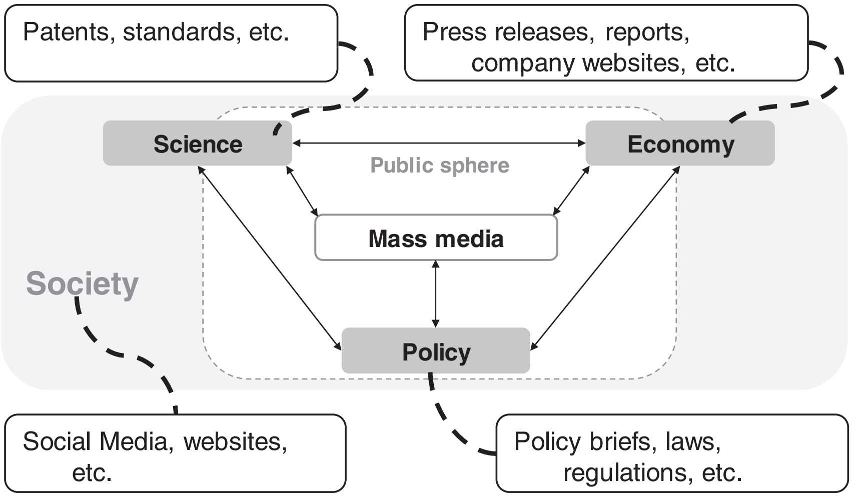
Patents, standards, etc.
Press releases, reports, company websites, etc.
Science
Economy
Public sphere
Mass media
Society
Policy
Social Media, websites, etc.
Policy briefs, laws, regulations, etc.

Fig. 5-1. Areas of innovation system and exemplary textual data sources.

economy, science) in more detail. Also the role of agency should then be discussed. However, the aim of this article was to develop a methodology to examine dynamics at the intersection of science, media and society rather than examining structural differences of innovation systems.

The results of this work 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 work. Finally, automatic approaches for a rapid overview of large datasets are relevant in our present time of increasing volume of data.

References

Alkemade, F., Kleinschmidt, C., Hekkert, M.P., 2007. Analysing emerging innovation systems: a functions approach to foresight. Int. J. Foresight Innov. Policy 3 (2), 139–168.

Andersen, A.D., Andersen, P.D., 2014. Innovation system foresight. Technol. Forecast. Soc. Chang. 88, 276–286.

Armbrust, M., Fox, A., Griffith, R., Joseph, A.D., et al., 2010. A view of cloud computing. Commun. ACM 53 (4), 50–58.

Balahur, A., Steinberger, R., 2009. 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.

Balahur, A., Steinberger, R., Kabadjov, M., Zavarella, V., et al., 2013. Sentiment analysis in the news. arXiv Preprint arXiv:1309.6202, pp. 2216–2220.

Bastian, M., Heymann, S., Jacomy, M., 2009. 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.

Batista, J., Ferreira, R., Tomaz, H., Ferreira, R., et al., 2015. A quantitative and qualitative assessment of automatic text summarization systems. In: Vanoirbeek, C., Genevès, P. (Eds.), The 2015 ACM Symposium, pp. 65–68 (Lausanne, Switzerland).

Bergek, A., Jacobsson, S., Carlsson, B., Lindmark, S., Rickne, A., 2008. Analyzing the functional dynamics of technological innovation systems: a scheme of analysis. Res. Policy 37 (3), 407–429.

Bird, S., Klein, E., Loper, E., 2009. Natural Language Processing with Python. first ed. O'Reilly, Beijing, Cambridge [Mass.].

Burkart, R., 2002. Kommunikationswissenschaft: Grundlagen und Problemfelder. fourth ed. Böhlau, Wien [et al.].

Cagnin, C., Amanatidou, E., Keenan, M., 2012. Orienting European innovation systems towards grand challenges and the roles that FTA can play. Sci. Public Policy 39 (2), 140–152.

Cobo, M.J., López-Herrera, A.G., Herrera-Viedma, E., Herrera, F., 2011. An approach for detecting, quantifying, and visualizing the evolution of a research field: a practical application to the Fuzzy Sets Theory field. J. Informetrics 5 (1), 146–166.

Cooke, P., 2001. Regional innovation systems, clusters, and the knowledge economy. Ind. Corp. Chang. 10 (4), 945–974.

Cozzens, S., Gatchair, S., Kang, J., Kim, K.-S., et al., 2010. Emerging technologies: quantitative identification and measurement. Tech. Anal. Strat. Manag. 22 (3), 361–376.

Cunningham, S.W., Porter, A.L., Newman, N.C., 2006. Special issue on tech mining: tech mining: exploiting science and technology information resources. Technol. Forecast. Soc. Chang. 73 (8), 915–922.

Daim, T.U., Rueda, G., Martin, H., Gerdsri, P., 2006. Forecasting emerging technologies: use of bibliometrics and patent analysis. Tech. Forcasting Soc. Chang. 73 (8), 981–1012.

Das, D., Martins, A.F., 2007. A Survey on Automatic Text Summarization.

Deacon, D., 2007. Yesterday's papers and today's technology: digital newspaper archives and 'push button' content analysis. Eur. J. Commun. 22 (1), 5–25.

Edquist, C. (Ed.), 1997. Systems of Innovation: Technologies, Institutions, and Organizations. Pinter, London.

Franzen, M., Rödder, S., Weingart, P., Maasen, S., Kaiser, M., Reinhart, M., Sutter, B., 2012. Wissenschaft und Massenmedien: Von Popularisierung zu Medialisierung. Handbuch Wissenschaftssoziologie. Springer Fachmedien Wiesbaden, Wiesbaden, pp. 355–364.

Freeman, C., 1987. Technology, Policy, and Economic Performance: Lessons from Japan. Pinter Publishers, London, New York.

Fuller, T., Loogma, K., 2009. Constructing futures: a social constructionist perspective on foresight methodology. Futures 41 (2), 71–79.

Glänzel, W., 2012. Bibliometric methods for detecting and analysing emerging research topics. Profesional Informacion 21 (2), 194–201.

Glenisson, P., Glänzel, W., Janssens, F., de Moor, B., 2005. Combining full text and bibliometric information in mapping scientific disciplines. Inf. Process. Manag. 41 (6), 1548–1572.

Grimmer, J., Stewart, B.M., 2013. Text as data: the promise and pitfalls of automatic content analysis methods for political texts. Polit. Anal. 1–31.

Hekkert, M.P., Negro, S.O., 2009. Functions of innovation systems as a framework to understand sustainable technological change: Empirical evidence for earlier claims. Technol. Forecast. Soc. Chang. 76 (4), 584–594.

Hekkert, M.P., Suurs, R.A.A., Negro, S.O., Kuhlmann, S., Smits, R., 2007. Functions of innovation systems: a new approach for analysing technological change. Technol. Forecast. Soc. Chang. 74 (4), 413–432.

Holz, F., Teresniak, S., 2010. Towards automatic detection and tracking of topic change. In: Gelbukh, A., Gelbukh, A. (Eds.), Proceedings of the 11th International Conference on Computational Linguistics and Intelligent Text Processing. Springer, pp. 327–339.

House, R.L., Iha, N.Y.M., Coppo, R.L., Alibabaei, L., et al., 2015. Artificial photosynthesis: where are we now? Where can we go?: where are we now? Where can we go? J Photochem Photobiol C: Photochem Rev.

Kabalak, A., Priddat, B.P., Rhomberg, M., 2008. Medien als Schnittstelle zwischen politischen und ökonomischen Strukturen - Politische Kommunikation in der Perspektive der Institutionenökonomie. In: Pfetsch, B., Adam, S. (Eds.), Massenmedien als politische Akteure. VS Verlag für Sozialwissenschaften, Wiesbaden, pp. 52–70.

Karnowski, V., 2013. Diffusionstheorie. In: Schweiger, W., Fahr, A. (Eds.), Handbuch Medienwirkungsforschung. Springer Fachmedien Wiesbaden, Wiesbaden, pp. 513–528.

Kostoff, R.N., 2012. Text mining for science and technology - a review part I – characterization/scientometrics. Scientometrics 1 (1), 11–21.

Krippendorff, K., 2013. Content Analysis: An Introduction to its Methodology. third ed. Los Angeles, London, SAGE.

Leydesdorff, L., Milojević, S., 2015. Scientometrics. In: Wright, J.D. (Ed.), International Encyclopedia of the Social & Behavioral Sciences, second ed. Elsevier, Amsterdam, pp. 322–327.

Leydesdorff, L., Welbers, K., 2011. The semantic mapping of words and co-words in contexts. J. Informetrics 5 (3), 469–475.

Li, N., Wu, D.D., 2010. Using text mining and sentiment analysis for online forums hotspot detection and forecast. Decis. Support. Syst. 48 (2), 354–368.

Luhmann, N., 2009. Die Realität der Massenmedien. fourth ed. VS, Verlag für Sozialwissenschaften, Wiesbaden.

Malerba, F., 2002. Sectoral systems of innovation and production: innovation systems. Res. Policy 31 (2), 247–264.

Martin, B.R., Johnston, R., 1999. Technology foresight for wiring up the national innovation system. Technol. Forecast. Soc. Chang. 60 (1), 37–54.

Montoyo, A., Martínez-Barco, P., Balahur, A., 2012. Subjectivity and sentiment analysis: an overview of the current state of the area and envisaged developments. Decis. Support. Syst. 53 (4), 675–679.

Moore, G.A., 2006. Crossing the Chasm: Marketing and Selling Disruptive Products to Mainstream Customers. Collins Business Essentials, New York, NY.

Negro, S.O., 2007. Dynamics of technological innovation systems: the case of biomass energy. Neth. Geogr. Stud. 356.

O'Connor, B., Banmann, D., 2011. Computational text analysis for social science: model assumptions and complexity. Public Health 41 (42), 1–7.

Öner, M.A., 2010. On theory building in foresight and futures studies: a discussion note. Futures 42 (9), 1019–1030.

Pollak, S., Coesemans, R., Daelemans, W., Lavra, N., 2011. Detecting contrasting patterns in newspaper articles by combining discourse analysis and text mining. Pragmatics 647–683.

Ravi, K., Ravi, V., 2015. A survey on opinion mining and sentiment analysis: tasks, approaches and applications. Knowl.-Based Syst. 89, 14–46.

Rogers, E.M., 1995. Diffusion of Innovations. fifth ed. Free Press, New York.

Scharkow, M., 2012. Automatische Inhaltsanalyse und maschinelles Lernen. epubli GmbH, Berlin.

Schenk, M., 2012. Medienwirkungsforschung. third ed. Mohr Siebeck, Tübingen.

Sculley, D., Pasanek, B.M., 2008. Meaning and mining: the impact of implicit assumptions in data mining for the humanities. Lit. Linguist. Comput. 23 (4), 409–424.

Stauffacher, M., Muggli, N., Scolobig, A., Moser, C., 2015. Framing deep geothermal energy in mass media: the case of Switzerland. Technol. Forecast. Soc. Chang. 98, 60–70.

Tigabu, A.D., Berkhout, F., van Beukering, P., 2015. The diffusion of a renewable energy technology and innovation system functioning: comparing bio-digestion in Kenya and Rwanda. Technol. Forecast. Soc. Chang. 90, 331–345.

van Eck, N.J., Waltman, L., Noyons, E.C.M., Buter, R.K., 2010. Automatic term identification for bibliometric mapping. Scientometrics 82 (3), 581–596.

Waldherr, A., 2008. Gatekeeper, Diskursproduzenten und Agenda-Setter — Akteursrollen von Massenmedien in Innovationsprozessen. In: Pfetsch, B., Adam, S. (Eds.), Massenmedien Als Politische Akteure. VS Verlag für Sozialwissenschaften, Wiesbaden, pp. 171–195.

Waldherr, A., 2012. The mass media as actors in innovation systems. In: Bauer, J., Lang, A., Schneider, V. (Eds.), Innovation Policy and Governance in High-Tech Industries. Springer, Berlin, Heidelberg, pp. 77–100.

Weaver, D.A., Bimber, B., 2008. Finding news stories: a comparison of searches using lexisnexis and google news. J. Mass Commun. Q. 85 (3), 515–530.

Victoria Kayser has been working as a researcher at Fraunhofer ISI in the Competence Center Foresight since April 2012. Prior to this, she studied Information Engineering and Management at the Karlsruhe Institute of Technology (KIT). Her current research concentrates on the integration of text mining in foresight what is spread out in her doctoral thesis at the TU Berlin.