Journal of Business Economics manuscript No.
(will be inserted by the editor)

A Systematic Literature Review of Mining Weak Signals and Trends for Corporate Foresight

Christian Mühlroth · **Michael Grottke**

Accepted: February 2018

Abstract Due to the ever-growing amount of data, computer-aided methods and systems to detect weak signals and trends for corporate foresight are in increasing demand. To this day, many papers on this topic have been published. However, research so far has only dealt with specific aspects, but it has failed to provide a comprehensive overview of the research domain. In this paper, we conduct a systematic literature review to organize existing insights and knowledge. The 91 relevant papers, published between 1997 and 2017, are analyzed for their distribution over time and research outlets. Classifying them by their distinct properties, we study the data sources exploited and the data mining techniques applied. We also consider 8 different purposes of analysis, namely weak signals and trends concerning political, economic, social and technological factors. The results of our systematic review show that the research domain has indeed been attracting growing attention over time. Furthermore, we observe a great variety of data mining and visualization techniques, and present insights on the efficacy and effectiveness of the data mining techniques applied. Our results reveal that a stronger emphasis on search strategies, data quality and automation is required to greatly reduce the human actor bias in the early stages of the corporate foresight process, thus supporting human experts more effectively in later stages such as strategic decision making and implementation. Moreover, systems for detecting weak signals and trends need to be able to learn and accumulate knowledge over time, attaining a holistic view on weak signals and trends, and incorporating multiple source types to provide a solid foundation for strategic decision making. The findings presented in this paper point to future research opportunities, and they can help practitioners decide which sources to exploit and which data mining techniques to apply when trying to detect weak signals and trends.

Keywords Machine Learning · Weak Signal Detection · Emerging Trend Detection · Corporate Foresight · Environmental Scanning · Strategic Decision Making · Big Data

JEL Classifications C8 · C88 · E17 · M1 · M19

Michael Grottke
Chair of Statistics and Econometrics, Friedrich-Alexander-Universität Erlangen-Nürnberg
Tel.: +49-911-5302 290
Fax: +49-911-5302 276
E-mail: michael.grottke@fau.de
This is a post-peer-review, pre-copyedit version of an article published in the Journal of Business Economics.
The final authenticated version is available online at: https://dx.doi.org/10.1007/s11573-018-0898-4

1 Introduction

The space for strategic decision making and managerial action is becoming increasingly complex. Companies need to navigate and act in an environment that is subject to constant change and growing managerial options. This challenge is amplified by incomplete and asynchronous information and the accelerating confrontation with new and unknown factors (Rohrbeck and Bade, 2012). For companies, these dynamics may provide new opportunities for growth (Veugelers et al, 2010), but they can also represent a fundamental threat (von der Gracht et al, 2010). Therefore, companies are searching for strategic orientation whilst being challenged to maintain their competitive advantage or even their existence.

Recently, the evolution of big data has been enabling companies to detect relevant weak signals and trends in their corporate environment by accessing and analyzing large amounts of data. The growing data sets used by environmental scanning contain information on political, economic, social and technological developments (Hiltunen, 2008). Since the amount of data is increasing too fast for humans to keep pace, scalable computer-aided systems are needed to extract and exploit it for creating insights and knowledge about the changing corporate environment (Keller and von der Gracht, 2014).

Research on corporate foresight has emphasized the need to detect changes in the corporate environment as early as possible (Ansoff, 1975; Rohrbeck et al, 2015). The changes to be identified range from an unexpected discontinuity to the possible emergence of a (mega-)trend which can fundamentally change technology and society (Ansoff, 1975; Kuosa, 2010). To delineate the terminology, Saritas and Smith (2011) defined weak signals as "early signs of possible but not confirmed changes that may later become more significant indicators of critical forces" and trends as "change factors that arise from broadly generalizable change and innovation". Weak signals and trends are usually considered the start and the end point, respectively, in the development from the earliest signal to a widespread trend (Hiltunen, 2008; Kuosa, 2010; Saritas and Smith, 2011); however, the exact development paths proposed by the various authors differ.

Research has already focused on various aspects of these challenges. Kontostathis et al (2004) reviewed different approaches and software tools for identifying emerging technology trends and news from the internet using text mining techniques. Taking the perspective of information systems research, Mayer et al (2011) evaluated environmental scanning systems and made suggestions for improving the future applicability of computer-aided systems. Rohrbeck and Bade (2012) provided a review and conceptual integration of environmental scanning, futures research, strategic foresight, and organizational future orientation; the authors tried to clearly distinguish between the terminology used in these research streams. Al-Azmi (2013) described various data, text and web mining techniques as well as practical case studies to assist business intelligence in terms of gaining competitive advantage, improved customer relationship management and fraud detection. Abbas et al (2014) reviewed existing literature on detecting technological trends in patent databases, whereas Eckhoff et al (2014) analyzed tools and methods for identifying emerging trends from large data collections (e.g., weak signals from web sites or emerging terms in conference publication abstracts). Blomqvist (2014) specifically researched the usage of semantic web technologies for business decision support and provided additional insights from expert interviews on this research stream. Furthermore, Wanner et al (2014) published a state-of-the-art report on visualization techniques that are applied for event detection in text mining. Madani (2015) employed network and cluster analysis using the CiteSpace software tool to analyze the literature on technology mining.

While previous research has already investigated specific aspects of this research domain, the following systematic literature review aims at holistically organizing insights and knowledge. We try to gather the prior research on exploiting big data for providing companies with valuable and future-relevant insights about dynamics and developments in their corporate environment. Based on the literature collected, we systematically analyze the development of detecting weak signals and trends in big data sets over time, and we provide an overview of the state-of-the-art methods. Our investigation sheds light on various research areas that are interconnected with weak signal detection and emerging trend detection, it identifies research gaps and points to future research directions.

The remainder of this article is structured as follows. In Section 2, we develop our research questions, while Section 3 describes the method employed for conducting the systematic literature review. The results are presented in Section 4 and are used as a basis for our discussion of findings in Section 5, pointing out the research gaps identified in our review and proposing aspects that future research should focus on. Finally, Section 6 concludes the paper.

2 Research questions

The postulation of research questions that drive the entire methodology of a systematic literature review is of critical importance (Kitchenham and Charters, 2007). In the following paragraphs, we therefore develop 6 research questions that will be answered in the course of our work.

Reviews on detecting weak signals (Ansoff, 1975; Eckhoff et al, 2014) and emerging trends (Kontostathis et al, 2004) in the corporate environment expected a future growth of these research disciplines. Recent research on corporate foresight emphasized the ever-increasing need for companies to analyze their corporate environment in an age of accelerating competition and constant change (Mayer et al, 2011; Rohrbeck and Bade, 2012). With the advances in computing power, computer-aided systems for detecting weak signals and trends are in demand (Keller and von der Gracht, 2014). It can therefore be assumed that the entire research domain consisting of these two research disciplines has been attracting growing interest in recent years. Moreover, existing research on related research disciplines, such as environmental scanning systems in general (Mayer et al, 2011), found that the relevant literature had been published in various journals and conference proceedings. To check the distribution of the publications in our research domain both in terms of time and in terms of the journals and conferences where they appeared, we formulate the following research question:

RQ 1. *How are the publications in the research domain distributed over time and outlets?*

Existing research revealed that various data sources had been used. For example, Abbas et al (2014) reviewed methods for detecting technological trends in patent data, whereas Eckhoff et al (2014) analyzed methods for discovering weak signals and trends based on scientific literature or web data. To develop a better understanding of the types of data sources used, we suggest the following research question:

RQ 2. *Which types of data sources are exploited for the detection of weak signals and trends for corporate foresight?*

Previous research on detecting weak signals and trends from big data collections is related to various purposes. For example, Kontostathis et al (2004) studied existing literature on emerging trend detection with text mining. The authors identified different fields of application, including detecting technological opportunities and changes in news events. A literature review conducted by Abbas et al (2014) analyzed the detection of technological trends and related domains (e.g., novelty detection and strategic technology planning) dealing with patent data. Eckhoff et al (2014) reviewed various approaches for detecting weak signals and emerging trends with computer-aided methods, such as discovering emerging terms in conference abstracts. Based on this prior work, it can be assumed that there are various purposes of analysis. To systematize both the literature and the existing approaches, the following research question is proposed:

RQ 3. *Which purposes of analysis are addressed by the existing research on detecting weak signals and trends for corporate foresight?*

Even in the early years of this research domain, methods and tools to assist human experts in extracting and analyzing knowledge from the ever-growing amount of data were considered to gain in importance (Fayyad et al, 1996). Naturally, human individuals are not anymore capable of coping with the large data volumes (Abbas et al, 2014). Recent research pointed out that work performed by human experts should shift from the early stages of data mining (such as data collection) to the later stages of data interpretation and decision making (Keller and von der Gracht, 2014). Again, this implies the need for efficient computer-aided systems. To understand and systematize the data mining techniques appropriate for mining weak signals and trends, the following research question is proposed:

RQ 4. *Which data mining techniques are used to detect weak signals and trends for corporate foresight?*

The main goal of data mining tasks is to process and visualize data in a way that facilitates interpretation by human experts (Steinecke et al, 2011), resulting in knowledge for decision making (Fayyad et al, 1996) and implementation strategies for corporate foresight (Keller and von der Gracht, 2014). To gain an understanding of how the data mining results can be interpreted, we formulate the following research question:

RQ 5. *How are results from mining weak signals and trends for corporate foresight interpreted?*

The results from mining weak signals and trends for corporate foresight are intended to support strategic decision making (Keller and von der Gracht, 2014). Therefore, the evaluation of newly introduced methods and algorithms plays a critical role when assessing their efficacy and effectiveness for various purposes. Moreover, Kontostathis et al (2004) reviewed systems for detecting emerging trends and pointed out that the selection of the performance metrics used is critical for the evaluation results. By answering the following research question, we seek for evidence on the efficacy and effectiveness of the methods and algorithms employed in the literature:

RQ 6. *How are efficacy and effectiveness of the approaches to detecting weak signals and trends for corporate foresight evaluated?*

3 Method

According to the structured frameworks proposed by Kitchenham and Charters (2007) and vom Brocke et al (2015), a systematic review consists of four steps: research identification and search, selection and quality assessment, data extraction and classification, and data synthesis. To answer the research questions formulated in the last section, we have performed a systematic literature review following this process; as discussed in the following subsections, we have adapted it to our research domain.

3.1 Research identification and search

To identify the most relevant literature, we searched the Web of Science (WoS) database. WoS includes all journals that are listed in the Science Citation Index Expanded (SCI-EXPANDED), and it thus provides a sufficiently broad range of literature.

A set of keywords was developed using an iterative refinement approach, weighing feasibility against coverage (vom Brocke et al, 2015). At first, general keywords such as "weak signal*" or "trend*" in combination with "(data OR text) AND mining" were used to search the WoS database. The set of keywords then was refined iteratively using high-frequency keywords from the literature obtained. Moreover, we distilled additional keywords based on the title and abstract of the literature found using the structured keyword stemming approach by Ferber (2003). The final query is shown in Table 1.

As for the publication dates, papers that were published before the early work on the weak signal theory by Ansoff (1975) were not considered in the analysis because our research aimed at discovering the state of the art in the research domain rather than tracing back its entire development path (Rohrbeck and Bade, 2012). The latest article included was published in March 2017. In total, 935 results were obtained using the database query.

Table 1 Search query and result count

Query	TI=((”weak signal*” OR trend* OR technolog* OR topic* OR ”research f*” OR ”technolog* opportunit*” OR converg* OR fusion) AND (converg* OR detect* OR discov* OR emerg* OR evol* OR identif* OR mining OR monitor* OR scan* OR trac*)) AND TS=(((trend OR data OR text) AND mining) OR ”trend detection” OR ”technolog* forecast*” OR ”technolog* intelligence” OR ”technolog* opportunit*” OR ”emerging topic*” OR ”topic detection” OR ”topic tracking”)
Result Count	935

3.2 Selection and quality assessment

The selection of search keywords was deliberately chosen to cover a broad spectrum of possible keyword combinations. Consequently, the result set retrieved from WoS was expected to contain a large number of papers not addressing the intended research domain. For this reason, it was necessary to further filter down the search results.

Each paper was reviewed manually, by reading its title and abstract. Papers that either did not address our research domain or did not use data mining techniques were removed. For example, the papers entitled "Coal in West Virginia: Geology and Current Mining Trends" (Gayle and Blake, 1980) and "Mining trends and patterns of software vulnerabilities" (Murtaza et al, 2016) had been received by the query. Both papers do not refer to the intended research domain, although their titles contain the terms "trend" and "mining". Overall, 888 papers were excluded for such reasons.

Another 5 papers were removed based on their full text. For example, the paper "Trend mining in social networks: from trend identification to visualization" (Nohuddin et al, 2014), which had initially been considered as relevant for our research domain, turned out to suggest a social network trend mining framework for identifying trends in cattle movements throughout Great Britain. Since this paper was not referring to corporate foresight, it was excluded.

Based on the entire selection and quality assessment process, a total of 893 papers were excluded, leaving us with 42 papers from the initial WoS search results.

We then carried out a forward and backward search to discover further relevant documents that had not been returned by the structured database query (vom Brocke et al, 2015). 49 additional journal and conference papers were selected. Thus, all in all 91 papers were included in the systematic literature review. For these papers, we conducted a detailed content analysis. Figure 1 shows the steps in a flowchart.

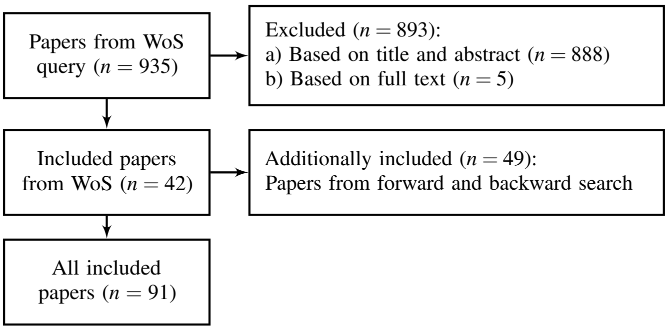

Fig. 1 Flowchart of the selection and quality assessment step, including forward and backward search

3.3 Data extraction and classification

Data from the papers included was extracted in a systematic way for 4 different aspects: research discipline, data source type, purpose of analysis, and data mining technique.

First, papers were classified by their *research discipline*. As mentioned before, research on corporate foresight has emphasized the need for detecting both weak signals (as the earliest possible detection of discontinuities) and emerging trends in the corporate environment (Ansoff, 1975; Kuosa, 2010). Thus, we classified all papers according to their focus by distinguishing between the two research disciplines *weak signal detection* and *emerging trend detection*. We differentiated the papers obtained based on their individual intentions.

For example, the papers by Thorleuchter et al (2014) ("Semantic weak signal tracing") and Ena et al (2016) ("A methodology for technology trend monitoring: the case of semantic technologies") clearly formulate their objectives; they were classified as belonging to the research discipline weak signal detection and emerging trend detection, respectively. Papers that did not explicitly state their goal in the title or abstract, such as the paper by Yoon and Park (2007) ("Development of new technology forecasting algorithm: Hybrid approach for morphology analysis and conjoint analysis of patent information") were classified based on the intended results of the data mining process. In these cases, we have distinguished between detecting early signals, single changes or anomalies (weak signal detection) and monitoring medium to long-term developments over time (emerging trend detection).

Second, the existing research was grouped by the exploited *data source types*. We developed the data source types described below in Section 4.2 using an inductive coding approach (Bernard, 2006). The papers were then classified based on these categories.

Third, the *purpose of analysis* was another aspect studied. We built the categories used based on the PEST classification from environmental scanning research, sometimes also referred to as the STEP classification (Hiltunen, 2008; Carr and Nanni, 2009). This classification scheme distinguishes between political, economic, social and technological (including scientific research and development) factors. Combining it with our two research disciplines, we defined a total of 8 purposes an analysis can address, from detecting political weak signals to identifying technological trends. The categorization thus allows a deeper analysis than the one based on the research discipline alone, and it will be used to structure our discussion in the course of this systematic literature review.

Fourth, three aspects concerning the employed *data mining techniques* were analyzed.

For one, the data mining techniques were classified according to their *data mining approach*. Vidhya and Aghila (2010) and Al-Azmi (2013) mainly distinguished between text mining (analyzing patterns within unstructured text data that needs to be structured in advance), bibliometric analysis (analyzing patterns within structured data, e.g., meta-data such as citations or pre-structured databases) and web mining (analyzing patterns within unstructured text data from the internet). For this research, we included web mining in the category text mining, because it also addresses the task of analyzing unstructured text data. As overlaps in text and data mining approaches can occur, the following three disjoint classes were used for categorizing the papers: *text mining*, *bibliometric analysis* and *joint text mining and bibliometric analysis*.

Next, the analysis of the data mining techniques was structured along the *data mining process* as proposed by Fayyad et al (1996). More specifically, we considered the 5 process steps data collection, data cleaning and pre-processing, data projection and transformation, data mining, and data visualization.

Finally, we specifically analyzed the employed *data mining methods* to provide a detailed view of possible patterns and methodological trends in combination with other categories. The data mining methods were grouped together according to the data mining task classification suggested by Fayyad et al (1996):

1. *Change and Deviation Detection*: identification of unusual data records or temporal patterns;
2. *Clustering*: discovery of groups and structures that are similar to a certain extent;
3. *Classification*: assignment of new data to existing categories;
4. *Dependency Modeling*: searching for relationships between variables, e.g., using rule-based approaches, pattern recognition, or network models;
5. *Regression*: finding a function which models the data with the least error;

6. *Summarization*: generating a more compact representation of the data.

3.4 Data synthesis

The research questions formulated in Section 2 will be answered in Section 4 using both quantitative and qualitative evidence from the systematic literature review. More detailed information, such as a complete list of the papers included, the categories assigned for some of the aspects studied, and tables with the complete classification results, are shown in the Appendix.

4 Results

The results presentation in this section is structured according to the research questions answered (see Section 2).

4.1 Distribution of the publications over time and outlets

In total, 91 papers were included in the systematic literature review. The earliest paper retrieved is from 1997. When we conducted our systematic literature review, 2 relevant papers published in 2017 were found. However, this does not allow any conclusions about recent developments, because the year 2017 was still in progress at that time. For this reason, we used all retrieved papers up to and including the year 2016 to study the development of the research domain. Figure 2 shows how the papers contained in our collection are distributed over time, distinguishing between the two research disciplines.

The first paper was published about two decades ago. Lent et al (1997) proposed a system to identify trends in US patents by analyzing sequential patterns employing shape queries of generalized sequential textual patterns over time. A few years later, Tho et al (2003) suggested a web mining system to collect scientific research publications for detecting technological trends in scientific research data.

Between 2005 and 2016, the research domain experienced an almost monotonic growth in terms of the number of publications, with notable outliers in the years 2006 and 2010. Overall, the results indicate that the research domain on detecting weak signals and trends for corporate foresight using data mining has attracted growing attention over the past few years. Our data further reveals that large parts of the growth observed are related to the research discipline emerging trend detection, whereas the number of papers in the research discipline weak signal detection remained almost constant.

The 91 papers analyzed were spread over a total of 52 different journals and conference proceedings. Most of them were published in Scientometrics (18 papers). This accumulation may be explained by the fact that this journal especially deals with quantitative features and characteristics of science and scientific research. In contrast, only 6 papers appeared in the journal at rank 2, Technological Forecasting and Social Change. While this journal also publishes research on the methodology and practice of technological forecasting and future studies, it does not predominantly focus on quantitative aspects, such as results and statistics from data mining. Rank 3, with 4 papers each, is shared by the Proceedings of the ACM International Conference on Knowledge Discovery and Data Mining, as well as the journals Expert Systems with Applications and PLoS ONE.

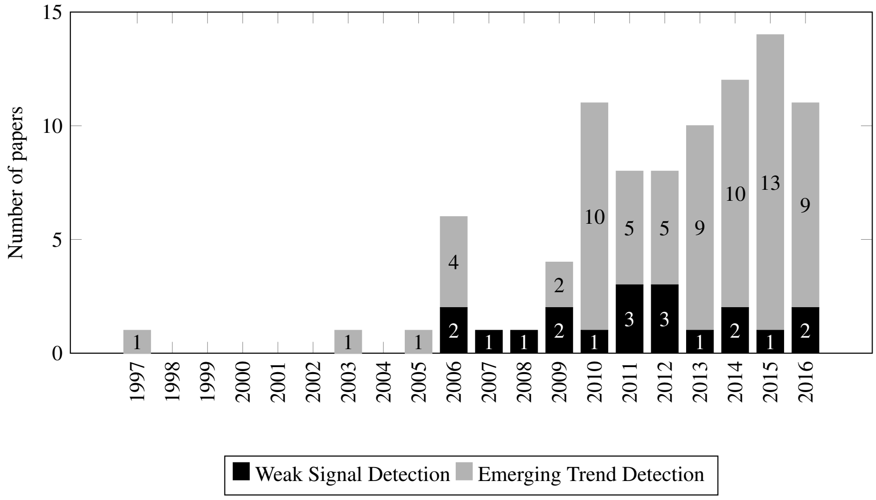

Fig. 2 Number of papers published between 1997 and 2016 for each research discipline

4.2 Exploited data source types

Overall, 5 categories of exploited data sources were found in the 91 papers analyzed. *Scientific publications* were used in 29 papers, *patents* were employed in 34 papers, *web sources* were exploited in 14 papers, and 23 papers made use of data from *social media*. Only 2 papers used an *other* data source type, as described below. It is to be noted that the total number of exploited data sources sums up to 102, which is greater than the total number of papers analyzed. The reason for this is that a few studies used more than one source type for their data analysis; this fact will be further analyzed in Section 4.3.

Scientific publications, including both journal and conference papers, were collected from WoS (Santo et al, 2006; Glänzel and Thijs, 2012; Wang et al, 2015b), Scopus (Woon et al, 2009), the National Digital Science Library of Korea (NDSL; Kim et al, 2012), and specific journals or magazines (Mei and Zhai, 2005; Wang and McCallum, 2006; Bolelli et al, 2009).

The specific sources for patent data included databases from the European Patent Office (EPO; Chang et al, 2010; Veugelers et al, 2010; Caviggioli, 2016), the World Intellectual Property Organization (WIPO; Veugelers et al, 2010), the Japan Patent Office (JPO; Veugelers et al, 2010), the Derwent World Patents Index (DWPI; Wu et al, 2010; Huang et al, 2015; Ma and Porter, 2015), the State Intellectual Property Office of the People's Republic of China (SIPO; Trappey et al, 2011), the United States Patent and Trademark Office (USPTO; Gerken and Moehrle, 2012; Jun et al, 2012a; Geum et al, 2013) and the NDSL (Kim et al, 2012), as well as the Australian Government Intellectual Property Department Database (IP Australia; Chen et al, 2015).

Among the web sources exploited were web news sites and newsgroups (Nasraoui et al, 2006; Liu et al, 2009; Kim et al, 2013), websites and weblogs (Goorha and Ungar, 2010; Veugelers et al, 2010; Thorleuchter et al, 2014), and RSS feeds (Dueñas-Fernández et al, 2014).

Moreover, data collected from the source type social media included social bookmarks from del.icio.us (Wetzker et al, 2010), topic-specific online communities and forums (Lu et al, 2013), Twitter (Adedoyin-Olowe et al, 2013; Rill et al, 2014; Song et al, 2014), Sina Weibo (a Chinese microblogging platform; Luo et al, 2015), Tencent QQ (a Chinese instant messaging platform; Wang et al, 2015a), Wikipedia (Kämpf et al, 2015), and image meta-data from Flickr (Bao et al, 2015).

As mentioned before, the data source type of 2 papers was classified as "other". First, the paper by Moreira et al (2015) used a pre-populated database of weak signals that had been populated manually by domain experts. This database contained any information that appeared to be relevant to these experts, formulated as summarizing sentences (e.g., "Wind energy may be a development factor in Brazil"). Second, the paper by Ena et al (2016) used specific databases containing foresight projects, European Commission projects and an online newspaper and magazine aggregator.

4.3 Purposes of analysis

As pointed out in Section 3.3, we distinguish between 8 different purposes of analysis.

Detecting *economic weak signals* aims at finding changes that refer to macro- and microeconomic changes in the corporate environment. These changes are expected to occur in areas of interest that are strategically relevant for companies concerning economic stability (Bun and Ishizuka, 2006) as well as surprising economical announcements and events (Liu et al, 2009). 2 papers were classified to pursue this purpose of analysis.

Mining data for *technological weak signals* focuses on the identification of early signals and changes in the (scientific) research and development (R&D) space. The main goal is not to identify and track technological developments over time, but rather to find the "needle in the haystack" in large document collections to support corporate decision making concerning R&D investments (Veugelers et al, 2010). Results from the analysis can be used to develop awareness for technological threats and opportunities or to support idea generation (Yoon and Park, 2007) and strategic technology planning (Yoon and Kim, 2012). Findings are also probed for their fit with existing internal technological competencies (Veugelers et al, 2010), and are evaluated for their practical relevance to support researchers, practitioners and companies as early as possible in deciding whether to pursue or neglect new developments (Geum et al, 2013). In total, 17 papers have been classified as belonging to this category.

Interestingly, none of the papers analyzed applied weak signal detection to mining for *political weak signals* and *social weak signals*. In fact, research in the domain dealing with social factors has postulated that the related data basis is perceived as a constant stream of text with unmanageable heterogeneity. Thus, if a system tried to detect weak signals in such data, it can be expected to produce a huge number of results, spamming its users (Goorha and Ungar, 2010; Bao et al, 2015; Pinto et al, 2015).

Relating to the shift from traditional communication channels (such as newspapers and TV) to online communication channels (such as websites, weblogs and social media), mining data for *political trends* can assist in detecting political topics and opinions (Rill et al, 2014; Song et al, 2014) as well as political news and events (Mei and Zhai, 2005; Gaul and Vincent, 2017) that are likely to influence the corporate environment. 4 of our papers have been classified as targeting this purpose of analysis.

Shifts in the entire economy are analyzed by detecting *economic trends* in publicly available data, related to both macro- and microeconomic factors. It includes corporate-related financial news (Dai et al, 2010) as well as the analysis of industry convergence patterns (Preschitschek et al, 2013; Weenen et al, 2013; Kim et al, 2015b). The latter category specifically observes blending boundaries between previously disjoint areas in markets or industries. This purpose of analysis has been identified in 5 papers.

The increasingly-available data concerning all societal aspects of human mankind is considered to serve as a sensor for real-world events (Aiello et al, 2013). Thus, the detection of *social trends* enables companies to react to new and emerging trends even faster than before (Cataldi et al, 2013). The applications are manifold: a company may quickly discover (shifts in) the customers' opinions about the company and its products (Goorha and Ungar, 2010; Luo et al, 2015). Furthermore, quickly-emerging health-related topics (Lu et al, 2013; Parker et al, 2013), hot product recommendations and incidences (Fang et al, 2014) and intensively-discussed general topics in society (Wang et al, 2015a) are subjects of mining social trends. In our literature corpus, we have found 22 papers belonging to this category.

Papers mining for *technological trends* aim at identifying the R&D landscape of a particular area, monitoring its developments over time (Ena et al, 2016). By quickly adapting to new technological developments, as well as by defending against possible threats and exploring new business opportunities, the analysis of research and technological trends is used to support corporate R&D investment decisions (Shih et al, 2010; Lee et al, 2011), thus encouraging companies to develop innovative products and technology strategies (Nguyen et al, 2016). While in the past experts from R&D tended to determine technology trends and investment decisions based on intrinsic knowledge and experience (Wang et al, 2010), modern approaches make use of complex data mining techniques (Woon et al, 2009; Chen et al, 2015; Nguyen et al, 2016). A specialized form of detecting trends in the technology space is the creation and usage of technology roadmaps. Its intention is to integrate the entire historical trajectory of a technology or a technology area, in combination with its current state of the art and knowledge that points towards future developments to predict future pathways (Huang et al, 2015; Jeong and Yoon, 2015; Ma and Porter, 2015; Wang et al, 2015b). Moreover, the convergence of entire technological areas can be identified by mining data for technological convergence patterns (Curran and Leker, 2011). Such patterns can be differentiated into supply-side convergence, i.e., the conversion of technological functionality, and demand-side convergence, including the satisfaction of latent needs based on the converged technological capabilities (Kim et al, 2014; Caviggioli, 2016). Overall, 41 of all papers obtained have been assigned to the purpose of mining technological trends.

Figure 3 visualizes the distribution of all papers over the various purposes of analysis.

Analyzing the purposes of analysis in conjunction with the exploited data source types from Section 4.2, our data naturally reveals clear associations between them. However, some remarkable outliers can be identified.

For the detection of technological weak signals, 14 out of 17 papers relied on either scientific publications or patents as the only data source type. One exception for this purpose of analysis is the semantic weak signal tracing approach by Thorleuchter et al (2014), using web data; another exception, employing a pre-populated database of weak signals manually collected by domain experts, was presented by Moreira et al (2015). Veugelers et al (2010) incorporated scientific publications, patents and web sources to provide a more holistic overview of the detected technological weak signals.

Technology trend identification made use of scientific publications and patent databases in all 41 cases. Hereby, the papers by Wu et al (2010), Curran and Leker (2011) and Kim et al (2012) exploited a combination of patents and scientific publications. Adding to this, Ena et al (2016) introduced a systematic monitoring approach to gain multi-perspective knowledge on technological trends based on scientific publications, patents, web sources, and other data such as foresight projects, European Commission projects and an online newspaper and magazine aggregator.

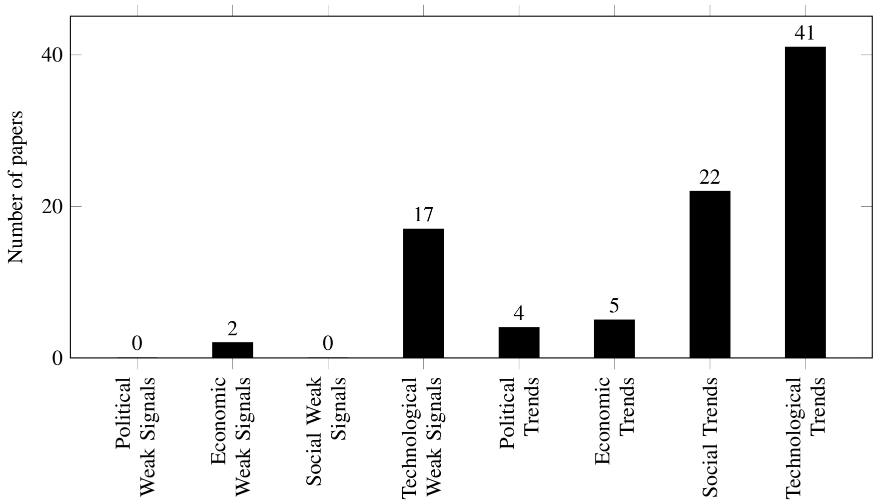

Fig. 3 Distribution of papers for each purpose of analysis

Economic trends pointing towards the convergence of entire markets and industries have been detected in patents (Preschitschek et al, 2013; Weenen et al, 2013) and web sources (Kim et al, 2015b), whereas other economic shifts were observed in financial news data (Dai et al, 2010) and social media data (Wetzker et al, 2010).

In summary, 83 of the 91 papers analyzed made use of one single data source type, revealing a clear association between the purposes of analysis and the data source types. However, 8 papers exploited more than 1 source type to reach their goals: 6 out of these 8 papers analyzed 2 different source types (Goorha and Ungar, 2010; Wu et al, 2010; Curran and Leker, 2011; Kim et al, 2012; Dueñas-Fernández et al, 2014; Bao et al, 2015), 1 paper analyzed 3 different source types (Veugelers et al, 2010), whereas Ena et al (2016) exploited 4 different source types in their approach. We did not find any papers analyzing data from more than 4 different source types.

Table 2 visualizes the distribution of purposes of analysis combined with the exploited data source types.

4.4 Data mining techniques

For answering the research question on the data mining techniques, we structure our results based on the data mining technique classes and subclasses formulated in Section 3.3. First, we present our results concerning the data mining approaches used. Second, we study the data mining techniques employed along the data mining process. Moreover, we specifically zoom into the data mining method perspective to learn more about possible commonalities and outliers. As none of the papers in our corpus aimed at mining political and social weak signals, those two categories are neglected in the further course of this analysis.

Table 2 Data source types exploited for the different research disciplines and purposes of analysis

		Weak Signal Detection		Emerging Trend Detection	
Single source type	Scientific Publications	Technological Weak Signals	6	Technological Trends	18
	Patents	Technological Weak Signals	8	Economic Trends Technological Trends	2 19
	Web Sources	Economic Weak Signals Technological Weak Signals	2 1	Political Trends Economic Trends Social Trends	2 2 2
	Social Media	n/a	-	Political Trends Economic Trends Social Trends	2 1 17
	Other	Technological Weak Signals	1	n/a	-
	Total		**18**		**65**
Multiple source types	Scientific Publications + Patents	n/a	-	Technological Trends	3
	Web Sources + Social Media	n/a	-	Social Trends	3
	Scientific Publications + Patents + Web Sources	Technological Weak Signals	1	n/a	-
	Scientific Publications + Patents + Web Sources + Other	n/a	-	Technological Trends	1
	Total		**1**		**7**

4.4.1 Data mining approach

We first analyzed all 91 papers for their data mining approaches, taking into account their research disciplines and purposes of analysis. A more detailed analysis of the individual data mining methods will be presented in Section 4.4.5. Table 3 summarizes the absolute frequencies, abbreviating weak signal detection by WSD, and emerging trend detection by ETD.

Out of all papers analyzed, text mining (58 papers) was applied far more often than bibliometric analysis (18 papers) and joint analysis (15 papers). Shifting the perspective to the combination of data mining approaches and purposes of analysis, bibliometric analysis as the single data mining approach was not used for detecting economic weak signals and political trends. Moreover, a joint application of text mining and bibliometric analysis was primarily applied in technological domains, namely technological trends (12 papers) and technological weak signals (2 papers).

It is noteworthy that purposes of analysis which focus on identifying political, economic and social factors tend to primarily rely on text mining, whereas purposes of analysis which focus on detecting weak signals and trends in the R&D space make use of both text mining and bibliometric analysis. The only exception to this observation is the paper by Liu et al (2009), applying a joint analysis on web news pages to detect economic weak signals. This observation could indicate that these approaches have been particularly promising for the respective purposes of analysis.

Table 3 Research disciplines, purposes of analysis, and data mining approaches

		Text Mining	Bibliometric Analysis	Joint Analysis	**Total**
WSD	Economic Weak Signals	1	0	1	**2**
	Technological Weak Signals	11	4	2	**17**
	Total WSD	**12**	**4**	**3**	**19**
ETD	Political Trends	4	0	0	**4**
	Economic Trends	4	1	0	**5**
	Social Trends	20	2	0	**22**
	Technological Trends	18	11	12	**41**
	Total ETD	**46**	**14**	**12**	**72**
	Total Overall	**58**	**18**	**15**	**91**

4.4.2 Data collection

In data mining, queries are used to limit down the retrieved data in order to focus on specific contents (Abbas et al, 2014; Thorleuchter et al, 2014). If no queries are used, too much data is collected, thus increasing the probability of irrelevant results (Goorha and Ungar, 2010). With a query that is too broad, results may be too general. However, if the query is too narrow, there is a risk of missing important data. For queries created by human experts, results may be incomplete or biased towards the experts' field of expertise (Grandjean et al, 2005; Milanez et al, 2014; Huang et al, 2015), thus introducing human bias into the data collection process (Palomino et al, 2013).

Out of all 91 papers analyzed, 30 papers did not use expert-based queries to collect the data from their targeted data sources, while 61 papers did. However, only 1 of these 61 papers applied a structured approach for generating the query: Milanez et al (2014) employed the modularized Boolean search strategy by Porter et al (2008) to retrieve patents from DWPI. All other 60 papers did not show any indication of a structured query creation and expansion. Instead, their queries were manually created to the best of the authors' or the involved experts' knowledge. We also did not see the application of any other approach to support query creation, such as automated query generation and expansion (Robertson, 2004; Imran and Sharan, 2010).

Figure 4 visualizes the distribution of expert query usage for each purpose of analysis.

Obviously, purposes of analysis exploiting economic and social factors are relying on expert-based queries to a lesser degree than purposes of analysis analyzing political and technological factors. One reason might be that research in the latter areas tries to answer specific questions in specific knowledge domains, like political trends concerning specific political parties (Rill et al, 2014), or specialized R&D topics like nano* and nanocrystal* technologies (Santo et al, 2006), magnetic random access memory technologies (Wang et al, 2010), or RFID technologies (Trappey et al, 2011).

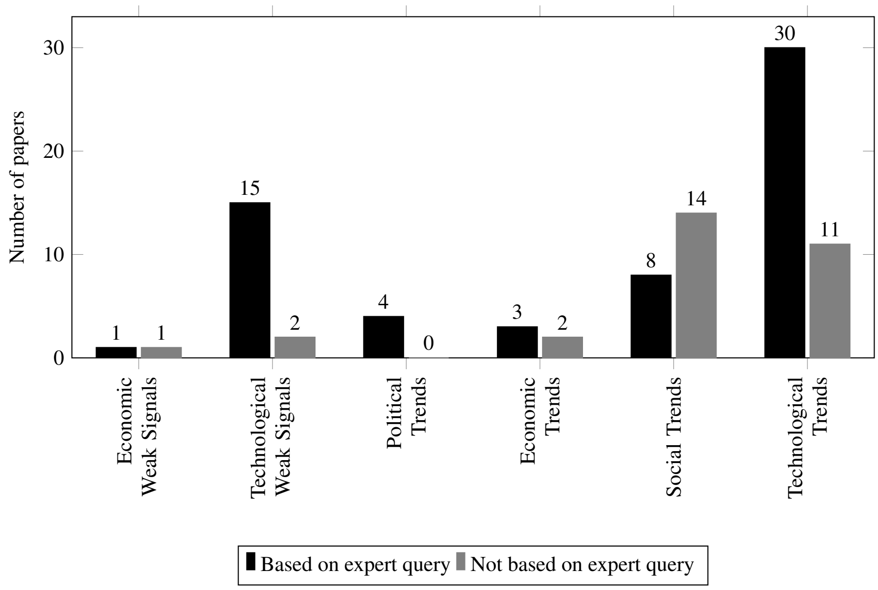

Fig. 4 Distribution of expert query usage for each purpose of analysis

Furthermore, most of the data collection techniques did not allow for corpus updates (76 of the 91 papers). This means that after collecting the data, the document collection was considered to be static; no further documents could be included anymore. If additional documents were found at a later point in time, then an entirely new analysis would have to be conducted, instead of refining the previous results based on the recently-detected documents.

Figure 5 depicts the use of data corpus updates for each purpose of analysis.

It is remarkable that hardly any approach in the technological domain allows for corpus updates. The sole exception is the one presented by Thorleuchter et al (2014), exploiting web data to detect technological weak signals. In contrast to this, the possiblity of corpus updates is much more wide-spread among approaches mining for economic, political, and social factors, which primarily focus on web sources and social media data (see Table 2). Indeed, these are also the data source types perceived as a constant stream of text contents by the research on social trends (Goorha and Ungar, 2010; Bao et al, 2015; Pinto et al, 2015) mentioned before in Section 4.3. Such information needs to be collected at a higher frequency than data from other sources, which explains why the related research often allows for corpus updates.

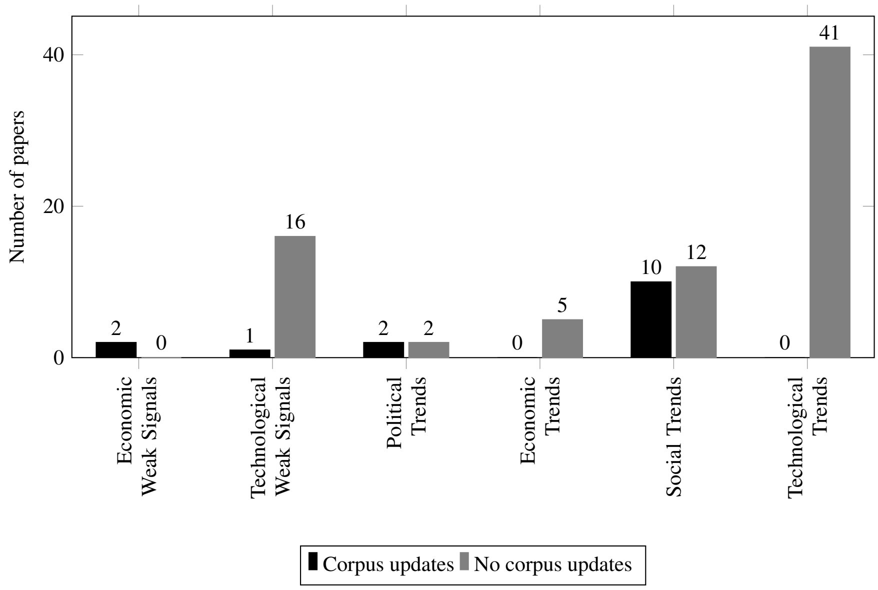

Fig. 5 Distribution of data corpus updates for each purpose of analysis

4.4.3 Data cleaning and pre-processing

The application of data cleaning and pre-processing techniques is expected to increase the performance of the data analysis (Aiello et al, 2013; Thorleuchter and van den Poel, 2013). 49 of the 91 papers made use of at least one data cleaning and pre-processing technique, while 8 papers explicitly stated that cleaning and pre-processing were not employed. In 34 cases, such techniques were not mentioned at all; this may hint at the fact that they were not performed, but it is impossible to say with certainty.

Techniques that were used include filtering of unwanted characters, punctuation or white-spaces (e.g., Bello-Orgaz et al, 2014; Luo et al, 2015; Wang et al, 2015a), case conversion to lowercase (e.g., Wang and McCallum, 2006; Guo et al, 2011; Thorleuchter et al, 2014), stemming (e.g., Wang et al, 2010; Lee et al, 2011; Lu et al, 2013), stop-word removal (e.g., Wang et al, 2010; Yoon and Kim, 2012; Luo et al, 2015), correction of typographical errors (Thorleuchter et al, 2014), feature filtering based on Zipf's law or variations (e.g., Wang and McCallum, 2006; Thorleuchter et al, 2014; Huang et al, 2015), feature filtering based on too little or too many characters (e.g., Wang and McCallum, 2006; Thorleuchter et al, 2014), fuzzy word matching (Huang et al, 2015), synonym replacement/aggregation (e.g., Aiello et al, 2013; Moreira et al, 2015) and expert-based involvement in cleaning and pre-processing the data (e.g., Guo et al, 2011; Kim et al, 2013; Huang et al, 2015).

Authors who did not make use of any data cleaning and pre-processing techniques gave various reasons for their choice. For example, Mei and Zhai (2005) intentionally did not apply any cleaning and pre-processing techniques to test the robustness of their newly-introduced algorithms. Cataldi et al (2013) wanted to consider all words that Twitter users submitted to update their status; these authors believed that noisy data may be filtered out by adapting standard text analysis methods using techniques similar to inverse frequency. Lee et al (2014) replaced the automatic data cleaning and pre-processing by iterative and manual tasks to generate and select keywords for further processing. Takahashi et al (2014) used Twittermentioning behaviors of users to detect anomalies in links that were then tested against their business for detecting emerging topics, where no pre-processing was required. Some authors, such as Shibata et al (2011), did not give any reasons for skipping the data cleaning and pre-processing step. In their discussion, Woon and Madnick (2012) commented that filtering and data cleaning mechanisms are required due to the inconsistent quality of data received from publicly-available databases, yet the authors did not apply any such technique to their data set. Jun et al (2012b) stated that they employed pre-processing techniques, but they in fact equated the task of pre-processing data with text mining and vector space projection of keywords into a document-term matrix; they did not perform any data cleaning and pre-processing task in the proper sense. Figure 6 summarizes the quantitative findings separately for each purpose of analysis.

We analyzed the applied techniques from other viewpoints as well, but we did not find any noticeable patterns. For example, 4 of the 8 papers that definitely did not apply any data cleaning and pre-processing techniques were concerned with text mining, while the other 4 applied bibliometric analysis. Although one might have expected in advance that it is mainly bibliometric data for which cleaning and pre-processing are not required, this does not seem to be the case.

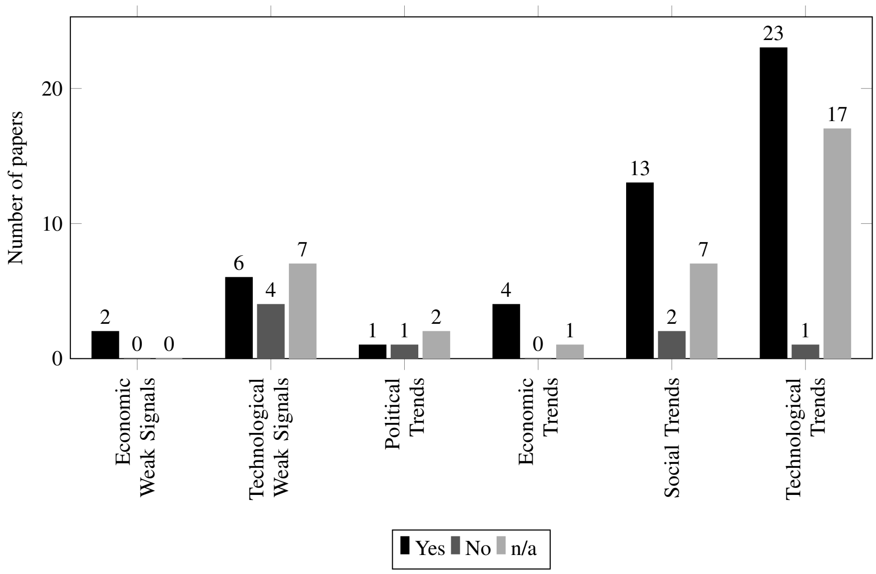

Fig. 6 Distribution of (non-)usage of data cleaning and pre-processing techniques for each purpose of analysis

4.4.4 Data projection and transformation

During the analysis of the literature, we found that there are two main approaches for performing data projection and transformation: automated and expert-assisted (manual) approaches. Automated approaches to project and transform the data into a structured format do not need any human involvement, whereas expert-assisted approaches require assistance in at least one step. A paper was classified to be expert-assisted if any human involvement in that particular task was observed, thus indicating that this task was not performed fully automated. Still, these papers might also apply automated techniques during some part of the process.

64 of all 91 papers employed completely automated data projection and transformation techniques, while 27 papers relied on human experts in at least in one step. Figure 7 depicts the distribution of automated and expert-assisted data projection and transformation techniques for each purpose of analysis.

It is remarkable that the detection of technological weak signals predominantly uses expert-assisted data projection and transformation techniques (9 out of 17 papers), whereas the other 5 purposes of analysis more frequently rely on fully-automated approaches. For example, Yoon and Park (2007) based the central part of their analysis (morphology matrix) on a predefined structure that was created iteratively by domain experts. Yoon and Kim (2012) made use of additional human screening in the process of pre-processing as well as screening meaningful subject-action-object (SAO) structures for further data mining. Geum et al (2013) described keyword extraction as a repetitive process in which expert judgment plays an important role in defining the keywords, and Lee et al (2014) applied an iterative, expert-assisted approach to create, select and expand the set of SAO structures.

Automated approaches included statistical, linguistic and semantic techniques from Natural Language Processing (NLP), such as Generalized Sequential Patterns (GSP; Lent et al, 1997), Named Entity Recognition (NER; Liu et al, 2009; Goorha and Ungar, 2010), Point-of-Speech (PoS) tagging (e.g., Abe and Tsumoto, 2010; Wang et al, 2015b; Nguyen et al, 2016), SAO structures (e.g., Gerken and Moehrle, 2012; Lee et al, 2014) and Latent Semantic Analysis (LSI; Thorleuchter et al, 2014), amongst others. When using text mining, these techniques were mainly employed to tag and filter extracted features for projecting the previously-unstructured text data into a structured vector space model, to which subsequent mining techniques could be applied.

Manual techniques included the creation of features on a completely manual basis (Lee, 2008), or at least human expert screening of automatically-generated keywords or features (Abe and Tsumoto, 2010; Lee et al, 2011; Yoon and Kim, 2012). Similar to the findings on expert assistance in creating queries for data collection from Section 4.4.2, any expert-assisted involvement in data projection and transformation may introduce bias into the data (Grandjean et al, 2005; Milanez et al, 2014; Huang et al, 2015).

4.4.5 Data mining methods

To systematize the existing literature with respect to the data mining methods, we used the classification scheme introduced in Section 3.3. More specifically, we employed a combined perspective of purposes of analysis and data mining methods to analyze the data towards possible patterns and methodological trends. Since the data mining process can involve multiple iterations and loops between any two mining steps performed (Fayyad et al, 1996), several data mining methods may be employed in one paper. We therefore classified each paper based on the data mining method that was predominantly used during the analysis. For example, Song et al (2014) based their analysis of the political landscape in Twitter data on probabilistic topic modeling using multinomial latent Dirichlet allocation (LDA), before plotting the topics identified as a time series to visualize which topics are rising over time. We thus determined classification, which includes LDA, to be the predominant data mining method in this paper. The detailed overview of the applied methods can be found in Table 5 in the Appendix.

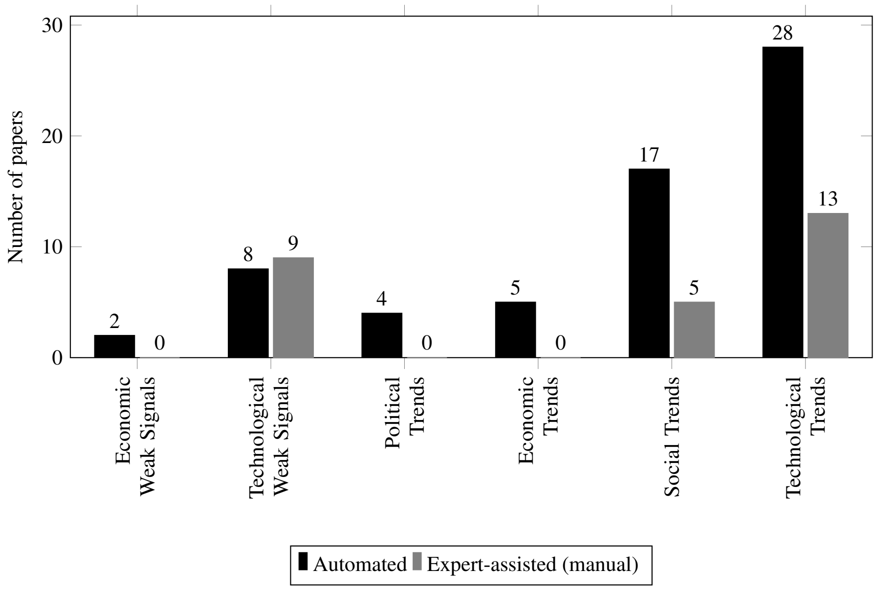

Fig. 7 Distribution of automated vs. expert-assisted data projection and transformation techniques for each purpose of analysis

The analysis of the literature revealed some interesting co-occurrences of data mining methods, purposes of analysis, and data source types that can be interpreted as common patterns and methodological trends.

First, our data showed that dependency modeling methods, such as network analysis and rule-based approaches, were mainly used in emerging trend detection, whereas weak signal detection barely made use of such methods: the detection of technological trends (41 out of 91 papers) predominantly relied on dependency modeling (17 papers), followed by change and deviation detection (9 papers) and clustering (7 papers) whereas the detection of technological weak signals (17 out of 91 papers) primarily focused on change and deviation detection as well as clustering methods (8 papers each). This may indicate that other methods are not very effective in detecting weak signals (or have not been applied so far), but are preferably used in such settings.

Next, it was found from the detailed analysis of the applied methods (see Table 5 in the Appendix) that classification approaches using probabilistic topic modeling such as LDA and other probabilistic models were predominantly applied when mining social trends (8 papers, e.g., Luo et al, 2015; Wang et al, 2015a; Yang et al, 2016; Xie et al, 2016) whereas only 2 papers used this method in modeling trends in R&D (Wang and McCallum, 2006; Bolelli et al, 2009). As mentioned in Section 4.3, the data used for detecting social trends is interpreted as a constant stream of textual data, which necessitates the possibility of corpus updates (see Section 4.4.2). The probabilistic topic models used are capable of saving and retaining their state of the model, thus satisfying this requirement.

Another common pattern revealed is that change and deviation detection, including time series and outlier analysis (26 out of 91 papers), was primarily found to be employed in mining technological weak signals and trends (17 papers, e.g., Shibata et al, 2011; Woon and Madnick, 2012; Nguyen et al, 2016), which mainly used scientific publications and patents as data source type but exploited web sources and social media data to a notably lesser degree.

From a timing perspective, it can be observed that research papers have recently started to apply more complex data mining methods. For example, detecting technological trends originally focused mainly on change and deviation detection and clustering methods; since 2014, there has been a shift to more complex algorithms (Huang and Chang, 2014; Wang et al, 2014; Cheng et al, 2015; Mryglod et al, 2016; Tu and Hsu, 2016). This trend is in line with the emerging number of applications of more sophisticated topic models (such as LDA and its extensions) when detecting social trends. Our observation may hint at the fact that the research domain has been maturing over time, now employing more complex models than when it initially emerged.

The quantitative findings of the analysis are summarized in Table 4, and a more detailed overview of the qualitative findings is given in the following paragraphs.

Table 4 Research disciplines, purposes of analysis, and data mining methods

		Change and Deviation Detection	Clustering	Classification	Dependency Modeling	Regression	Summarization	**Total**
WSD	Economic Weak Signals	0	0	0	1	0	1	**2**
	Technological Weak Signals	8	8	0	1	0	0	**17**
	Total WSD	**8**	**8**	**0**	**2**	**0**	**1**	**19**
ETD	Political Trends	1	2	1	0	0	0	**4**
	Economic Trends	2	1	1	1	0	0	**5**
	Social Trends	6	5	8	2	1	0	**22**
	Technological Trends	9	7	5	17	3	0	**41**
	Total ETD	**18**	**15**	**15**	**20**	**4**	**0**	**72**
	Total Overall	**26**	**23**	**15**	**22**	**4**	**1**	**91**

Despite the common patterns and trends observed, the data reveals a high heterogeneity of data mining methods applied. However, when grouping the data by the purpose of analysis, like in Table 5 in the Appendix, rough tendencies can be worked out.

When detecting technological weak signals, the main methods applied to patent data were based on change and deviation detection such as co-word analysis (Lee, 2008), term growth analysis (Guo et al, 2011), citation analysis (Shibata et al, 2011), and co-citation analysis (González-Alcaide et al, 2016). In contrast, methods that were predominantly applied to scientific publication data showed a certain tendency towards clustering techniques such as k-means (Veugelers et al, 2010), support-vector clustering (Jun et al, 2012b) or k-medoids clustering (Moreira et al, 2015) and the semantic SAO analysis (Gerken and Moehrle, 2012;

Yoon and Kim, 2012; Lee et al, 2014). Mining economic weak signals did no reveal a tendency towards commonly-used methods.

Identifying social trends predominantly applied probabilistic topic modeling such as LDA and modified topic models (Lau, 2012; Aiello et al, 2013; Luo et al, 2015; Xie et al, 2016), change and deviation detection using term growth analysis (Cataldi et al, 2013; Parker et al, 2013; Kim et al, 2015a) and various clustering approaches (Bello-Orgaz et al, 2014; Fang et al, 2014; Bao et al, 2015).

Applied methods in mining technological trends showed a tendency to make use of dependency modeling including network analysis using cluster-based networks (Lee et al, 2010; Wang et al, 2014; Mryglod et al, 2016; Tu and Hsu, 2016), co-occurrence networks (Wang et al, 2010), SAO networks (Choi et al, 2011) and co-citation networks (Barirani et al, 2013) as well as change and deviation techniques such as term growth analysis (Santo et al, 2006; Fan and Chang, 2008; Abe and Tsumoto, 2010; Park et al, 2011; Tu and Seng, 2012) and rule-based analysis, such as sequential pattern mining (Lent et al, 1997) and association rule mining (Shih et al, 2010; Jun et al, 2012a). As a sub-category, the creation of technology roadmaps was found to be a highly manual process requiring substantial expert involvement and being supported by keyword clustering (Jeong and Yoon, 2015; Ma and Porter, 2015), patent-citation network analysis (Huang et al, 2015) and SAO-based analysis (Wang et al, 2015b). Analyzing technological convergence patterns primarily relied on co-occurrence analyses – such as co-classification (Curran and Leker, 2011) and co-classification of patent IPC codes (Caviggioli, 2016) – and dependency modeling in the form of patent-citation network analysis (Kim et al, 2014).

For political and economic trends, our data did not reveal any patterns.

Although no universally-valid processing chain can be observed, our data showed clear tendencies concerning the data mining methods used (and therefore expected to deliver promising results) for specific purposes of analysis. These insights can be used to select the best processing methods for different purposes and contexts of corporate foresight.

4.4.6 Data visualization

Many papers employed more than one visualization technique to present their results without relying on a predominant form. Thus, the total number of visualization techniques encountered (190) is greater than the total number of papers (91).

The 3 most-frequently-used kinds of visualization were basic charts, such as line charts, pie charts, bar charts, column charts or tables (78 of all 91 papers), time series or timeline visualization (52 papers) and network or graph visualizations (28 papers). Additional visualization techniques employed were topic maps (9 papers), roadmaps (4 papers), scatterplots (4 papers), clouds (3 papers), radars (3 papers), world maps (2 papers) and classification trees (1 paper).

A combined analysis of the data visualization techniques employed, the data mining approaches and the purposes of analysis did not reveal any noteworthy patterns. However, it is remarkable that the papers analyzed primarily used simple visualization techniques to present their analysis results.

4.5 Result interpretation

To understand the authors' approach to interpreting weak signals and trends, papers were classified into 2 categories: 53 of the 91 papers used some quantitative criterion or statistic to interpret their results (e.g., statistical trend indicators), while the remaining 38 papers interpreted their results describing the analysis output qualitatively, not relying on any quantitative measure. Figure 8 visualizes the findings on the interpretation of analysis results with respect to the different purposes of analysis.

To give some examples for an interpretation based on quantitative statistics, Hennig et al (2013) interpreted trends in social networks using the average slope and intercept values that classify if a term is trending or not, and Nguyen et al (2016) interpreted technology trends as normalized occurrence weights of a specific term in a certain number of documents in a particular time slot.

With respect to qualitative interpretation, Lee (2008) classified hubs in the generated knowledge map, which in turn were considered as R&D weak signals. Similarly, Huang and Chang (2014) labeled the clusters obtained from hierarchically clustering highly-cited articles as R&D trends.

An interesting approach to enhance the interpretability of results was suggested by Veugelers et al (2010). The authors created technology intelligence profiles that were connected to external data sources. Further data was pulled from external sources (such as the number of detected companies, key companies, recent Google alerts and internet keywords) to provide domain experts with additional context information.

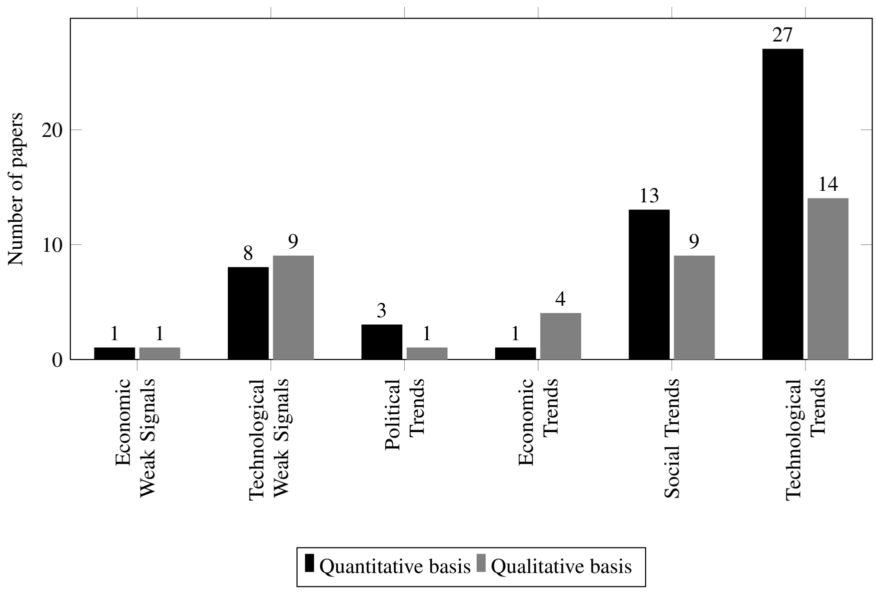

Fig. 8 Distribution of result interpretation on a quantitative and a qualitative basis for each purpose of analysis

In addition to the results interpretation, we observed an overlap in interpretation terminology in various papers. Research on detecting weak signals and trends tends to introduce a certain level of ambiguity of terminology. For example, automatically-identified topics (Lau, 2012) were labeled as trends by the authors. A literature overview by Eckhoff et al (2014) mixed the terminology on weak signals and trends when referring to “the application of a quantitative model for detecting weak signals (emerging trends) with the help of a[n] inference model and Bayesian network". To overcome this ambiguity, a study by Saritas and Smith (2011) provided a clear and concise distinction between the terminologies of future-related objects. Following these authors' research, weak signals are defined as "early signs of possible but not confirmed changes" that may represent first signs of future trends. In turn, trends are considered as "change factors that arise from broadly generalizable change and innovation" that last several years and that usually have a global reach. Furthermore, some authors used the term "topics", although they did not perform any typical topic modeling; they instead used this term to describe clusters generated by the usual clustering algorithms. For example, Tu and Seng (2012) stated that "an emerging trend is a topic area that is growing in interest and utility over time". Chang et al (2010) and Bao et al (2015) similarly mixed the terminology related to topics and clusters.

4.6 Evaluation techniques

Since one of the main goals of mining weak signals and trends is to increase the level of automation, the efficacy and effectiveness of newly-developed algorithms and proposed solutions need to be evaluated by using meaningful techniques.

In many of the papers analyzed, the authors applied multiple evaluation techniques. Thus, the total number of evaluation techniques applied (136) is greater than the total number of papers (91). Figure 9 shows the evaluation techniques used for each purpose of analysis.

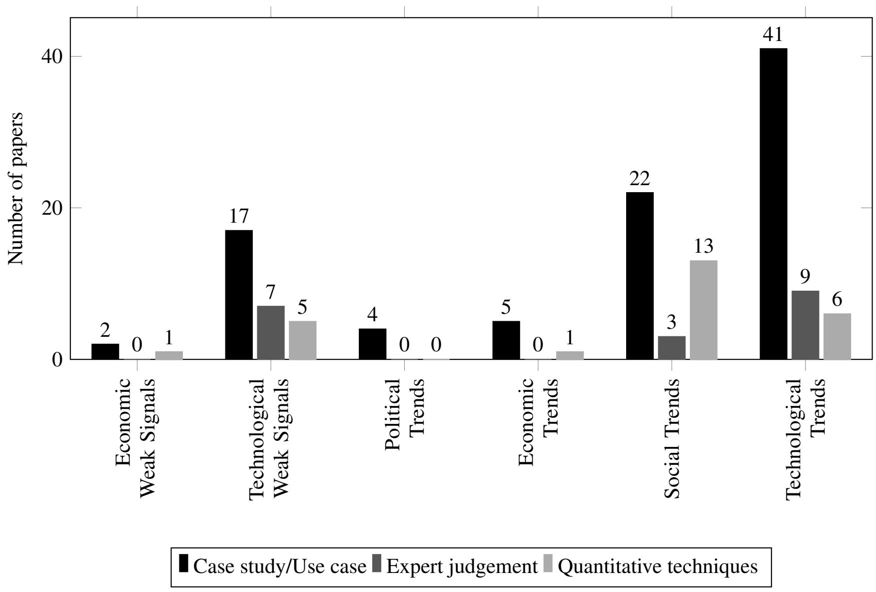

Fig. 9 Distribution of evaluation techniques for each purpose of analysis

All 91 papers employed use cases or case studies as a qualitative form of evaluation. With this technique, authors tested their methodology either on synthetic datasets (e.g., Chi et al, 2006; Takahashi et al, 2014; Bao et al, 2015) or on real datasets (e.g., Thorleuchter and van den Poel, 2013; Rill et al, 2014; Wang et al, 2015b). Typically, papers using this form of evaluation came to the conclusion that the results are promising.

A qualitative evaluation by external domain experts (different from the authors themselves) was additionally applied in 19 papers. In these papers, experts were asked to evaluate the effectiveness of the employed data mining approaches to detect weak signals and trends. In general, domain experts provided positive feedback on the results of the applied data mining techniques to the respective goal and purpose of analysis at hand (Veugelers et al, 2010; Huang and Chang, 2014; Thorleuchter et al, 2014).

More specifically, different levels of effectiveness can be observed by analyzing the experts' statements: some authors have criticized the outcomes to be too general and high-level (Tu and Seng, 2012; Ma and Porter, 2015), whereas both papers applied simple term growth and keyword clustering approaches without allowing for a finer-grained analysis. At the basic level, experts stated that the outcomes of the data mining analysis can provide a complete overview of the knowledge domain analyzed (Huang and Chang, 2014; Ena et al, 2016), that the proposed solutions are able to find core roots of innovations (Chen, 2006) and that the results generally are in line with the domain experts' expertise (Glänzel and Thijs, 2012; Huang and Chang, 2014). The presented algorithms and solutions were also seen as capable of counteracting the human actor bias, providing more objective evidence for weak signals and trends (Shibata et al, 2011; Jun et al, 2012b; Geum et al, 2013). Moreover, few experts stated that the applied methods can save valuable time (Lee et al, 2014) and have a positive financial impact by greatly reducing costs (Shibata et al, 2011; Wang et al, 2015b). The experts also confirmed that the applied methods identified new weak signals and trends that would otherwise have gone unnoticed by them (Shibata et al, 2011; Geum et al, 2013; Lee et al, 2014).

On the quantitative side, 26 papers employed specific performance indicators to evaluate the efficacy of the approaches, such as precision, recall, and the F-measure (Lau, 2012; Thorleuchter et al, 2014; Tu and Hsu, 2016), accuracy (Kim et al, 2012; Geum et al, 2013; Kim et al, 2015a) as well as perplexity (Wang et al, 2015a). In most cases, the statistics derived had high values.

Overall, the evaluations of the proposed methods still rely on use cases for the major part. Moreover, there may be the danger of a publication bias: the quantitative measures employed could have been chosen deliberately to yield high values, because manuscripts presenting unfavorable results of the proposed methods may have a reduced chance of getting accepted for publication. Many authors also claimed to have analyzed their data in real time, yet they did not evaluate the algorithm performance in terms of time or computational cost. These facts make it impossible to confirm the quantitative efficacy of the applied data mining methods when detecting weak signals and trends. A more detailed empirical study on the efficacy of the applied methods should therefore investigate the performance of the proposed approaches based on quantitative evidence.

Combining the qualitative evaluations from domain experts with the quantitative evaluations by the authors, there are indications that both the efficacy and effectiveness of mining weak signals and trends using the presented data mining methods can be rated positively, but there is still room for improvements and further research.

5 Discussion

Based on the results of our systematic literature review, we have derived 5 findings that we will present in the following subsections.

5.1 Need for improved search strategies

Obviously, the quality of the collected data has a significant impact on the quality of the data mining results (Al-Azmi, 2013; Blomqvist, 2014). The literature analyzed revealed that there is a fine line between different forms of search strategies and queries for data collection. The broader the query, including the possiblity of applying no query at all, the higher the coverage, thus lowering the risk of missing important data; however, this incurs the risk of too general results (Goorha and Ungar, 2010; Al-Azmi, 2013). In contrast to that, the more narrow the query, the lower the coverage, thus increasing the danger of missing important data but lowering the risk of too general results (Palomino et al, 2013).

Shifting the focus to the goals of weak signal detection and emerging trend detection, another dimension needs to be considered: as weak signal detection aims at identifying discontinuities as early as possible, a weak signal scanning system should possibly process as much data as possible in order to detect fine-grained changes in it. Naturally, the data used should have the highest coverage possible (broad or no query), but this has been found to quickly spam users as a lot of noise is introduced by irrelevant data (Goorha and Ungar, 2010; Bao et al, 2015; Pinto et al, 2015). To cope with this issue, most of the proposed solutions in the area of weak signal detection in our literature corpus used expert queries to limit the data coverage (16 out of 19 papers, about 84%). Emerging trend detection, on the other hand, analyzes medium- to long-term developments in order to detect trends using quantitative statistics and thus bases the analysis on a broader data base than weak signal scanning. Consequently, proposed solutions in the papers analyzed made use of expert queries to a lesser degree (45 out of 72 papers, about 63%) because their approaches are found to be more robust due to their broader data base. Thus, the decision between the breadth and depth of the applied query is suspected to be a critical success factor for the efficacy and effectiveness of the proposed approaches.

However, although 61 of the 91 papers analyzed used some form of expert-based queries, only 1 of them was found to employ a structured query generation process (Milanez et al, 2014). In all other approaches the search queries for data collection were created to the best of the authors' or the involved experts' knowledge.

Existing approaches to improving the data quality, such as automated query generation and expansion (Robertson, 2004; Imran and Sharan, 2010) were not employed in the included literature (see Section 4.4.2). Other research domains with similar issues concerning data quality have already employed methods to cope with them. For example, in terms of the performed process steps systematic literature reviews share important similarities with data mining. Both start with developing an understanding of the goals and the formulation of research questions; data is then collected, pre-processed and structured for further analysis and interpretation in the context of specific questions (Fayyad et al, 1996; Kitchenham and Charters, 2007; vom Brocke et al, 2015). Typical problems, such as the publication bias (Kitchenham and Charters, 2007) or limited literature coverage (vom Brocke et al, 2015), have already been addressed by research on literature reviews. For example, Ferber (2003) proposed a structured keyword stemming approach to create and expand search queries. Following this line of thought, it may be assumed that already-employed techniques from other research domains could also be used in data mining. We therefore suggest the following finding:

Finding 1. *Methods and lessons learned from other research domains with similar problems and limitations should be transferred to the research domain of data mining, in order to improve search strategies, data quality and results.*

5.2 Lack of automation

The literature analyzed, and especially the more recent research, showed a demand for a higher degree of automation. For example, Lee et al (2011) suggested to further systematize and automate their proposed method, as it contained a considerable number of steps that needed to be performed manually. Keller and von der Gracht (2014) postulated that large parts of the foresight process should be automated, allowing expert involvement to be moved from the early phases of data collection and transformation to interpretation and decision making, where human knowledge cannot (yet) be replaced by an automated solution (Lee et al, 2009; Veugelers et al, 2010).

Despite the demand for more automation, a considerable amount of papers used various methods that still need human expert involvement. While this seems to be reasonable for the task of deriving knowledge and decisions from the analysis results, 61 of the 91 papers (approximately 67%) used manually-compiled expert queries for data collection (see Section 4.4.2), 27 papers (about 30%) made use of human expert assistance in projecting and transforming the data (see Section 4.4.4), and 38 papers (approximately 42%) predominantly based their result interpretation on qualitative rather than quantitative aspects (see Section 4.5).

At the same time, the involvement of human experts in the data mining process (e.g., for creating queries and projecting and transforming the data, see Section 4.4.2) has been criticized to introduce bias towards the respective experts' field of expertise (Grandjean et al, 2005; Milanez et al, 2014; Huang et al, 2015). However, existing approaches have already proposed solutions on how to counteract this human actor bias by providing more objective evidence for weak signals and trends (Shibata et al, 2011; Jun et al, 2012b; Geum et al, 2013). Addressing these insights, the following finding is proposed:

Finding 2. *Research efforts towards increasing the level of automation are required to reduce the risk of introducing human actor bias into the data mining process and to shift the focus of valuable human resources to the later stages of result interpretation, decision making and strategy implementation.*

5.3 Need for machine learning approaches allowing corpus updates

While there have been approaches allowing the initially-downloaded document collection to be augmented with new documents in web sources and social media data, this is not the case for the majority of the papers analyzed (see Section 4.4.2). If a method does not allow updates to the existing document corpus, then the detection of new documents (e.g., new patents) requires the data mining process to be started from scratch using the entire document collection. Thus, the analysis cannot be based on prior results and knowledge. In fact, the outcomes of the new analysis may be different from and difficult to compare with the previous results due to the many factors influencing the text and data mining algorithms (e.g., the number of clusters to be determined or the number of iterations during a clustering process). Moreover, the risk of expert bias may weigh in. It thus seems that approaches without corpus updates are designed as one-time efforts rather than continuous detection and monitoring systems.

It has further been observed that research on purposes of analysis exploiting web sources and social media data claimed that traditional text and data mining approaches are not applicable because the data volume to be processed is larger than in other application domains (Goorha and Ungar, 2010; Bao et al, 2015; Pinto et al, 2015). But even researchers from these other papers reasoned that the amount of data is increasing too fast for humans to manually analyze and interpret it (Glänzel and Thijs, 2012; Huang and Chang, 2014; Wang et al, 2014). It may therefore be worthwhile to transfer machine learning techniques from approaches exploiting web sources and social media data to the analysis of other data source types. We thus postulate the following finding:

Finding 3. *Machine learning techniques allowing corpus updates, currently employed in approaches that exploit web sources and social media data, should also be considered for the analysis of other data source types.*

5.4 Trends represent the rearview mirror

In the existing literature, weak signals are considered as the earliest possible signs of changes that may develop into stronger indicators of future discontinuities, developments or trends. Trends, in contrast, are defined as functions of time developing into a certain direction, which means that a sufficiently large data base is required for detecting a trend (Hiltunen, 2008; Kuosa, 2010; Saritas and Smith, 2011).

During our systematic literature review, we found that in most cases the goal of the identified approaches is to detect changes as early as possible, to assist human experts and decision makers in identifying and defending against possible threats as well as to explore potential opportunities (see Section 4.3).

Following the definition and terminology of the term “trend” in recent research, it is evident that trends are only capable of describing developments that have already been occurring for a longer period of time. Thus, strategic decisions and actions in the context of corporate foresight that are exclusively based on the analysis of emerging trends may already be considered too late (Schwarz, 2005). This means that the actual goal of quickly identifying and reacting to changes and developments in the corporate environment is missed.

Consequently, trends can be used to identify and describe directed movements in data sets to understand latest developments from the past until today. In general, results from detecting emerging trends were found to be useful by external expert judgement (see Section 4.6). However, Tu and Seng (2012) and Ma and Porter (2015) found that domain experts saw more value in deeper-level analyses than in the initially-discovered trends. The authors used this insight as an opportunity to generate sub-level results that contained more specific information than their high-level clusters.

Furthermore, extrapolating trends into the future is not recommended due to the risk of ignoring non-linear, overlapping and surprising developments such as wild cards, shocks and other discontinuities (Kuosa, 2010; Saritas and Smith, 2011; Rohrbeck and Bade, 2012). Considering weak signals and their development towards possible trends (Kuosa, 2010; Saritas and Smith, 2011), signals of greater strength (Hiltunen, 2008) are also found to occur too late for strategic decision making in corporate foresight (Thorleuchter and van den Poel, 2013).

While a certain ambiguity both in the terminology and the interpretation of weak signals and trends can be observed in the literature (see Section 4.5), and while there have already been some research efforts concerning the conceptual integration of weak signals and trends (Hiltunen, 2008; Saritas and Smith, 2011; Rohrbeck et al, 2015), a meaningful integration of weak signals and trends from a data-centric and statistical viewpoint is still lacking:

Finding 4. *Further research on the integration and joint usage of weak signals and trends from a data-centric and statistical perspective is needed in order to develop integrated systems for meaningful and early decision making.*

5.5 Need for harmonization of multiple source types

Our results from Section 4.3 show that there are multiple purposes of analysis and use cases of detecting weak signals and trends to help companies decide and act as early as possible. Naturally, the exploited data sources are clearly associated with purposes of analysis. For example, weak signals and trends for R&D are usually detected based on scientific publications and patents, while political, economic and social factors are typically mined from web sources and social media data. It was also found that most of the papers (83 out of 91) used only one source type (such as patents from either one or multiple patent databases) to detect weak signals and trends (see Table 2). Besides, there are few exceptions where the authors either carried out a separate analysis for each data source type or involved human experts to a large extent in selecting and generating the features (see Section 4.2).

It is remarkable that, although all papers analyzed share the same goal (i.e., detecting weak signals and trends), they almost exclusively rely on a single data source type. Possible findings from other data source types, which might be of critical importance, are completely neglected, and the data used thus only provides a rather limited excerpt of reality (Shibata et al, 2011). Recapitulating one of the core goals of corporate foresight, namely to observe and monitor all relevant developments in its corporate environment, it remains questionable whether approaches only exploiting one data source type qualify for comprehensive early weak signal and trend detection because strong constraints on data coverage are inherent to the proposed solutions.

A possible explanation for why most of the proposed methods do not make use of multiple source types is given by Wu et al (2010). Based on their approach to detecting science and technology trends for technology intelligence, these authors concluded that "science goes earlier than technology in a life cycle". The authors thus postulated that early signals and developments of an emerging trend are first found in scientific publications and only later in patent databases. This observation may suggest that a joint analysis of multiple source types poses the risk of temporal distortion; the various source types first need to be examined for their temporal dependencies before analyzing them jointly. Furthermore, Veugelers et al (2010) postulated that the various source types require different data mining strategies due to their different properties. This hypothesis is backed by the fact that in most countries there is a time delay of about 18 months between a patent application and its publication (Milanez et al, 2014).

These insights may indicate that multiple views on various data source types are needed to identify the very early weak signals, to track their development into emerging trends and to truly understand how changes in the corporate environment develop and diffuse over time. A computer-aided system that integrates all perspectives into one holistic database and platform may therefore satisfy the overarching goal of supporting corporate strategic decision making and implementation. Thus, we formulate the following finding:

Finding 5. *There is a need for research on how to integrate multiple source types at various points in time to provide a holistic perspective on relevant weak signals and trends.*

6 Conclusions

In this paper, we have performed groundwork towards analyzing and understanding multi-perspective mining of weak signals and trends for corporate foresight. By conducting a systematic literature review with a total of 91 relevant papers, our research offers greater rigor than action research provides. We systematically analyzed the landscape of this research domain by reviewing 20 years of scientific research, classifying and summarizing the existing research from different perspectives comparing its distinct properties, and providing an overview of the state-of-the-art methods as well as their trends and new directions.

Literature reviews face the risk of various limitations that need to be addressed early. One frequently-encountered issue is limited literature coverage (vom Brocke et al, 2015). Naturally, our research is limited to a restricted number of papers. To counteract this, we have based our literature search on WoS which contains one of the largest publication indexes (SCI-EXPANDED). The structured stemming method due to Ferber (2003) has been applied for search query creation and iterative expansion to broaden the sample space. Additional forward and backward search including both journal papers and conference proceedings was performed to ensure adequate coverage and saturation. Our review thus provides a solid foundation for discussion, helping other researchers to follow and join in the development of this research domain. Another concern might be related to the categorization of to the data source types. Since the application of an inductive coding approach can produce different results when being applied by different persons, the classification presented in this paper entail a certain level of subjectivity and may legtimately be a subject of debate.

Our results have indicated that this research domain is attracting growing attention. The main reasons for this development are the ever-growing complexity of the corporate environment for companies in combination with the accelerating rates of new competition and innovations, as well as rapid changes in customer behavior. These factors force companies to search for strategies to navigate under high uncertainty. Moreover, this paper provides practical implications on the outcomes and usefulness of the proposed approaches and presents an overview of which methods are used to address various purposes in different contexts. Our systematic literature review may also be useful for organizations which are considering to start or to extend their activities in this domain to support human experts more effectively in strategic decision making processes.

We have presented five findings summarizing important insights which can serve as opportunities for future research in order to mature this research domain. As data volumes continue to grow, computer-aided systems will experience an even higher demand. For companies to develop higher trust in computer-aided systems supporting their strategic decision making processes, the best possible quality of data for mining weak signals and trends needs to be ensured, requiring improved search strategies. Moreover, in order to allow domain experts to shift from early stages (i.e., data collection and processing) to later stages of the corporate foresight process (i.e., interpretation, decision making and implementation), a higher degree of automation needs to be achieved, thus greatly reducing the human actor bias. Computer-aided systems should also be able to learn and accumulate knowledge over time, building upon prior knowledge, and ensuring real weak signal and trend monitoring based on an extensive data basis. Integrated systems with a holistic understanding of the interactions between weak signals and trends are required, as the mere detection of trends will not satisfy the need of companies to react as early as possible. Towards developing these methods and systems, information from diverse source types will be needed in order to be able to truly understand the development and diffusion of changes in the corporate environment over time.

Acknowledgments

This study was sponsored by the German Federal Ministry of Education and Research, grant number 02K16C191.

Appendix

To make the tables in this appendix more readable, we will use the following abbreviations for the categories introduced in Section 3.3 wherever appropriate:

1. Research Discipline (RSD)
   (a) Weak Signal Detection (WSD)
   (b) Emerging Trend Detection (ETD)
2. Data Mining Approach (DMA)
   (a) Text Mining (TM)
   (b) Bibliometric Analysis (BA)
   (c) Joint Analysis (JA)
3. Data Mining Task (DMT), grouping the Data Mining Methods (DMM)
   (a) Change and Deviation Detection (CDD)
   (b) Clustering (CLU)
   (c) Classification (CLA)
   (d) Dependency Modeling (DEM)
   (e) Regression (REG)
   (f) Summarization (SUM)
4. Data Mining Process (DMP)
   (a) Data Collection (DAC)
   (b) Data Cleaning and Pre-Processing (DCP)
   (c) Data Projection and Transformation (DPT)
   (d) Data Mining (DAM)
   (e) Data Visualization (DAV)

Table 5 Overview of included papers and their classification

Author(s)	Year	DMA	DMT	DMM
Political Weak Signals				**Total: 0**
Economic Weak Signals				**Total: 2**
Bun and Ishizuka	2006	TM	SUM	Summarization
Liu et al	2009	JA	DEM	Association Rules
Social Weak Signals				**Total: 0**
Technological Weak Signals				**Total: 17**
Schult and Spiliopoulou	2006	TM	CLU	Bisecting k-Means
Yoon and Park	2007	TM	CDD	Citation Analysis
Lee	2008	TM	CDD	Co-Word Analysis
Lee et al	2009	TM	CLU	Principal Component Analysis (PCA)
Veugelers et al	2010	JA	CLU	k-Means Clustering
Glänzel and Thijs	2011	BA	CLU	Hybrid Clustering
Guo et al	2011	TM	CDD	Term Growth Analysis
Shibata et al	2011	BA	CDD	Citation Analysis
Gerken and Moehrle	2012	JA	CDD	Subject-Action-Object (SAO) Analysis
Jun et al	2012	TM	CLU	Support-Vector Clustering
Yoon and Kim	2012	TM	CDD	Subject-Action-Object (SAO) Analysis
Geum et al	2013	TM	CLU	Mixture Models
Lee et al	2014	TM	CDD	Subject-Action-Object (SAO) Analysis
Thorleuchter et al	2014	TM	CLU	Latent Semantic Indexing (LSI)
Moreira et al	2015	TM	CLU	k-Medoids
González-Alcaide et al	2016	BA	CDD	Co-Citation Analysis
Rodriguez et al	2016	BA	DEM	Patent-Citation Network Analysis
Political Trends				**Total: 4**
Mei and Zhai	2005	TM	CLU	Mixture Models
Rill et al	2014	TM	CDD	Term Growth Analysis
Song et al	2014	TM	CLA	Latent Dirichlet Allocation (LDA)
Gaul and Vincent	2017	TM	CLU	Hierarchical Clustering
Economic Trends				**Total: 5**
Dai et al	2010	TM	CLU	Hierarchical Clustering
Wetzker et al	2010	TM	CLA	Modified Probabilistic Models
Preschitschek et al	2013	TM	CDD	Semantic Similarity
Weenen et al	2013	BA	DEM	Patent-Citation Network Analysis
Kim et al	2015	TM	CDD	Co-Occurence Analysis
Social Trends				**Total: 22**
Chi et al	2006	TM	CLU	Singular Value Decomposition (SVD)
Goorha and Ungar	2010	TM	CDD	Co-Word Analysis
Lau	2012	TM	CLA	Latent Dirichlet Allocation (LDA)
Adedoyin-Olowe et al	2013	TM	DEM	Association Rules
Aiello et al	2013	TM	CLA	Latent Dirichlet Allocation (LDA)
Cataldi et al	2013	TM	CDD	Term Aging Model
Hennig et al	2013	TM	CDD	Term Growth Analysis
Lu et al	2013	TM	CLU	Expectation Maximization Clustering
Parker et al	2013	TM	CDD	Term Growth Analysis
Bello-Orgaz et al	2014	TM	CLU	k-Means Clustering
Dueñas-Fernández et al	2014	TM	CLA	Latent Dirichlet Allocation (LDA)
Fang et al	2014	TM	CLU	Spectral Clustering
Takahashi et al	2014	BA	CDD	Change-Point Analysis

Author(s)	Year	DMA	DMT	DMM
Bao et al	2015	TM	CLU	Co-Clustering
Kämpf et al	2015	BA	DEM	Context Networks
Kim et al	2015	TM	CDD	Term Growth Analysis
Luo et al	2015	TM	CLA	Latent Dirichlet Allocation (LDA)
Pinto et al	2015	TM	REG	Hawkes Processes
Wang et al	2015	TM	CLA	Latent Dirichlet Allocation (LDA)
Xie et al	2016	TM	CLA	Latent Dirichlet Allocation (LDA)
Yang et al	2016	TM	CLA	Latent Dirichlet Allocation (LDA)
Huang et al	2017	TM	CLA	Modified Probabilistic Models
Technological Trends				**Total: 41**
Lent et al	1997	TM	DEM	Sequential Pattern Mining
Tho et al	2003	TM	CLU	Multi-Clustering Technique
Chen	2006	JA	CDD	Co-Citation Analysis
Santo et al	2006	BA	CDD	Term Growth Analysis
Wang and McCallum	2006	TM	CLA	Modified Probabilistic Models
Bolelli et al	2009	TM	CLA	Latent Dirichlet Allocation (LDA)
Woon et al	2009	BA	CLU	Taxonomy Generation
Abe and Tsumoto	2010	TM	CDD	Term Growth Analysis
Chang et al	2010	JA	DEM	Network Analysis
Fan and Chang	2010	TM	CDD	Term Growth Analysis
Lee et al	2010	JA	DEM	Cluster-Based Networks
Shih et al	2010	BA	DEM	Association Rules
Wu et al	2010	JA	CLU	Keyword Clustering
Wang et al	2010	TM	DEM	Co-Occurence Networks
Choi et al	2011	TM	DEM	Subject-Action-Object (SAO) Networks
Curran and Leker	2011	BA	CLA	Co-Classification
Lee et al	2011	TM	CLU	Formal Concept Analysis (FCA)
Park et al	2011	TM	CDD	Term Growth Analysis
Trappey et al	2011	TM	REG	Logistic Growth Curves
Jun et al	2012	JA	DEM	Association Rules
Kim et al	2012	BA	CLA	Decision Trees
Tu and Seng	2012	TM	CDD	Term Growth Analysis
Woon and Madnick	2012	BA	CDD	Co-Occurence Analysis
Barirani et al	2013	JA	DEM	Co-Citation Network
Huang and Chang	2014	BA	DEM	Bibliographic Coupling
Kim et al	2014	TM	DEM	Patent-Citation Network Analysis
Milanez et al	2014	BA	REG	Logistic Growth Curves
Wang et al	2014	TM	DEM	Co-Word Networks
Chen et al	2015	TM	REG	Piecewise Linear Representation (PLR)
Cheng et al	2015	BA	DEM	(Rule-based) Anomaly Detection
Huang et al	2015	JA	DEM	Patent-Citation Network Analysis
Jeong and Yoon	2015	TM	CLU	Keyword Clustering
Ma and Porter	2015	JA	CLU	Keyword Clustering
Wang et al	2015	TM	CDD	Subject-Action-Object (SAO) Analysis
Caviggioli	2016	BA	CLA	Co-Classification
Ena et al	2016	TM	CLU	Hierarchical Clustering
Mryglod et al	2016	JA	DEM	Co-Authorship Network Analysis
Nguyen et al	2016	JA	CDD	Term Growth Analysis
Noh et al	2016	JA	DEM	Patent-Citation Network Analysis
Park et al	2016	BA	DEM	Network Analysis
Tu and Hsu	2016	JA	DEM	Citation-Network Analysis
Overall				**Total: 91**

Table 6 Absolute frequencies used in research question 4

		RQ 4																												
		Expert Query based	Not based on Expert Query	Corpus Updates	No Corpus Updates	Data Pre-Processing	No Data Pre-Processing	n/a	Automated Feature Generation	Manual Feature Generation	Text Mining	Bibliometric Analysis	Joint Analysis	Change and Deviation Detection	Clustering	Classification	Dependency Modeling	Regression	Summarization	Basic Charts	Time Series/ Timeline	Network/ Graph	Topic Maps	Other	Roadmap	Scatterplot	Cloud	Radar	Worldmap	Classification Tree
WSD	Political Weak Signals	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Economic Weak Signals	1	1	2	0	2	0	0	2	0	1	0	1	0	0	0	1	0	1	2	0	0	0	0	0	0	0	0	0	0
	Social Weak Signals	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Technological Weak Signals	15	2	1	16	6	4	7	8	9	11	4	2	8	8	0	1	0	0	16	6	4	3	1	0	0	1	1	0	0
	Total WSD	**16**	**3**	**3**	**16**	**8**	**4**	**7**	**10**	**9**	**12**	**4**	**3**	**8**	**8**	**0**	**2**	**0**	**1**	**18**	**6**	**4**	**3**	**1**	**0**	**0**	**1**	**1**	**0**	**0**
ETD	Political Trends	4	0	2	2	1	1	2	4	0	4	0	0	1	2	1	0	0	0	3	4	2	1	0	0	0	1	1	0	0
	Economic Trends	3	2	0	5	4	0	1	5	0	4	1	0	2	1	1	1	0	0	5	4	1	0	0	0	1	0	0	0	0
	Social Trends	8	14	10	12	13	2	7	17	5	20	2	0	6	5	8	2	1	0	17	13	4	0	0	0	2	0	0	1	0
	Technological Trends	30	11	0	41	23	1	17	28	13	18	11	12	9	7	5	17	3	0	35	25	17	5	5	4	1	1	1	1	1
	Total ETD	**45**	**27**	**12**	**60**	**41**	**4**	**27**	**54**	**18**	**46**	**14**	**12**	**18**	**15**	**15**	**20**	**4**	**0**	**60**	**46**	**24**	**6**	**5**	**4**	**4**	**2**	**2**	**2**	**1**
	Total Overall	**61**	**30**	**15**	**76**	**49**	**8**	**34**	**64**	**27**	**58**	**18**	**15**	**26**	**23**	**15**	**22**	**4**	**1**	**78**	**52**	**28**	**9**	**6**	**4**	**4**	**3**	**3**	**2**	**1**

Table 7 Absolute frequencies used in research questions 2, 5 and 6

		RQ 2														RQ 5		RQ 6		
		Sci. Publications	Patents	Web Sources	Social Media	Other	1 Source Type	Different Types (parallel)	Different Types (jointly)	2 Source Types (parallel)	2 Source Types (jointly)	3 Source Types (parallel)	3 Source Types (jointly)	>3 Source Types (parallel)	>3 Source Types (parallel)	Quantitative	Qualitative	Use Case/Case Study	Expert Judgement	Quantitative Techniques
WSD	Political Weak Signals	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Economic Weak Signals	0	0	2	0	0	2	0	0	0	0	0	0	0	0	1	1	2	0	1
	Social Weak Signals	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Technological Weak Signals	7	9	2	0	1	16	1	0	0	0	1	0	0	0	8	9	17	7	5
	Total WSD	**7**	**9**	**4**	**0**	**1**	**18**	**1**	**0**	**0**	**0**	**1**	**0**	**0**	**0**	**9**	**10**	**19**	**7**	**6**
ETD	Political Trends	0	0	2	2	0	4	0	0	0	0	0	0	0	0	3	1	4	0	0
	Economic Trends	0	2	2	1	0	5	0	0	0	0	0	0	0	0	1	4	5	0	1
	Social Trends	0	0	5	20	0	19	1	2	1	2	0	0	0	0	13	9	22	3	13
	Technological Trends	22	23	1	0	1	37	3	1	2	1	0	0	1	0	27	14	41	9	6
	Total ETD	**22**	**25**	**10**	**23**	**1**	**65**	**4**	**3**	**3**	**3**	**0**	**0**	**1**	**0**	**44**	**28**	**72**	**12**	**20**
	Total Overall	**29**	**34**	**14**	**23**	**2**	**83**	**5**	**3**	**3**	**3**	**1**	**0**	**1**	**0**	**53**	**38**	**91**	**19**	**26**

References

Abbas A, Zhang L, Khan SU (2014) A literature review on the state-of-the-art in patent analysis. World Patent Information 37:3–13, DOI 10.1016/j.wpi.2013.12.006

Abe H, Tsumoto S (2010) Trend detection from large text data. IEEE International Conference on Systems Man and Cybernetics (SMC) pp 310–315, DOI 10.1109/ICSMC.2010.5641682

Adedoyin-Olowe M, Gaber MM, Stahl F (2013) Trcm: A methodology for temporal analysis of evolving concepts in twitter. In: International Conference on Artificial Intelligence and Soft Computing, Springer Berlin Heidelberg, pp 135–145, DOI 10.1007/978-3-642-38610-7_13

Aiello LM, Petkos G, Martin C, Corney D, Papadopoulos S, Skraba R, Goker A, Kompatsiaris I, Jaimes A (2013) Sensing trending topics in twitter. IEEE Transactions on Multimedia 15(6):1268–1282, DOI 10.1109/TMM.2013.2265080

Al-Azmi AAR (2013) Data, text and web mining for business intelligence: A survey. International Journal of Data Mining & Knowledge Management Process 3(2):1–21, DOI 10.5121/ijdkp.2013.3201

Ansoff HI (1975) Managing strategic surprise by response to weak signals. California Management Review 18(2):21–33

Bao BK, Xu C, Min W, Hossain MS (2015) Cross-platform emerging topic detection and elaboration from multimedia streams. ACM Transactions on Multimedia Computing, Communications, and Applications 11(4):1–21, DOI 10.1145/2730889

Barirani A, Agard B, Beaudry C (2013) Discovering and assessing fields of expertise in nanomedicine: A patent co-citation network perspective. Scientometrics 94(3):1111–1136, DOI 10.1007/s11192-012-0891-6

Bello-Orgaz G, Menendez H, Okazaki S, Camacho D (2014) Combining social-based data mining techniques to extract collective trends from twitter. Malaysian Journal of Computer Science 27(2):95–111

Bernard HR (2006) Social research methods: Qualitative and quantitative approaches, reprinted edn. SAGE, Thousand Oaks, Calif.

Blomqvist E (2014) The use of semantic web technologies for decision support – a survey. Semantic Web 5(3):177–201, DOI 10.3233/SW-2012-0084

Bolelli L, Ertekin Ş, Giles CL (2009) Topic and trend detection in text collections using latent dirichlet allocation. Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval (ECIR '09) DOI 10.1007/978-3-642-00958-7_84

vom Brocke J, Simons A, Riemer K, Niehaves B, Plattfaut R, Cleven A (2015) Standing on the shoulders of giants: Challenges and recommendations of literature search in information systems research. Communications of the Association for Information Systems 37(1)

Bun KK, Ishizuka M (2006) Emerging topic tracking system in www. Knowledge-Based Systems 19(3):164–171, DOI 10.1016/j.knosys.2005.11.008

Carr LP, Nanni AJ (2009) Delivering Results: Managing What Matters. Springer New York

Cataldi M, Di Caro L, Schifanella C (2013) Personalized emerging topic detection based on a term aging model. ACM Transactions on Intelligent Systems and Technology (TIST) 5(1):1–27, DOI 10.1145/2542182.2542189

Caviggioli F (2016) Technology fusion: Identification and analysis of the drivers of technology convergence using patent data. Technovation 55–56:22–32, DOI 10.1016/j.technovation.2016.04.003

Chang PL, Wu CC, Leu HJ (2010) Using patent analyses to monitor the technological trends in an emerging field of technology: A case of carbon nanotube field emission display. Scientometrics 82(1):5–19, DOI 10.1007/s11192-009-0033-y

Chen C (2006) Citespace ii: Detecting and visualizing emerging trends and transient patterns in scientific literature. Journal of the American Society for Information Science and Technology 57(3):359–377, DOI 10.1002/asi.20317

Chen H, Zhang G, Zhu D, Lu J (2015) A patent time series processing component for technology intelligence by trend identification functionality. Neural Computing and Applications 26(2):345–353, DOI 10.1007/s00521-014-1616-y

Cheng Q, Lu X, Liu Z, Huang J (2015) Mining research trends with anomaly detection models: The case of social computing research. Scientometrics 103(2):453–469, DOI 10.1007/s11192-015-1559-9

Chi Y, Tseng BL, Tatemura J (2006) Eigen-trend: Trend analysis in the blogosphere based on singular value decompositions. In: Proceedings of the 15th ACM International Conference on Information and Knowledge Management, New York, pp 68–77

Choi S, Yoon J, Kim K, Lee JY, Kim CH (2011) Sao network analysis of patents for technology trends identification: A case study of polymer electrolyte membrane technology in proton exchange membrane fuel cells. Scientometrics 88(3):863–883, DOI 10.1007/s11192-011-0420-z

Curran CS, Leker J (2011) Patent indicators for monitoring convergence – examples from nff and ict. Technological Forecasting and Social Change 78(2):256–273, DOI 10.1016/j.techfore.2010.06.021

Dai XY, Chen QC, Wang XL, Xu J (2010) Online topic detection and tracking of financial news based on hierarchical clustering. In: 2010 International Conference on Machine Learning and Cybernetics (ICMLC), pp 3341–3346, DOI 10.1109/ICMLC.2010.5580677

Dueñas-Fernández R, Velásquez JD, L'Huillier G (2014) Detecting trends on the web: A multidisciplinary approach. Information Fusion 20:129–135, DOI 10.1016/j.inffus.2014.01.006

Eckhoff R, Markus M, Lassnig M, Schön S (2014) Detecting weak signals with technologies: Overview of current technology-enhanced approaches for the detection of weak signals. International Journal of Trends In Economics, Management & Technology (IJTEMT) USA III(V):1–7

Ena O, Mikova N, Saritas O, Sokolova A (2016) A methodology for technology trend monitoring: The case of semantic technologies. Scientometrics 108(3):1013–1041, DOI 10.1007/s11192-016-2024-0

Fan TK, Chang CH (2008) Exploring evolutionary technical trends from academic research papers. In: Kise K (ed) The Eighth IAPR International Workshop on Document Analysis Systems, 2008, IEEE, Piscataway, NJ, pp 574–581, DOI 10.1109/DAS.2008.25

Fang Y, Zhang H, Ye Y, Li X (2014) Detecting hot topics from twitter: A multiview approach. Journal of Information Science pp 578–593, DOI 10.1177/0165551514541614

Fayyad U, Piatetsky-Shapiro G, Smyth P (1996) From data mining to knowledge discovery in databases. AI Magazine 17(3):37–54

Ferber R (2003) Information Retrieval: Suchmodelle und Data-Mining-Verfahren für Textsammlungen und das Web, 1st edn. Dpunkt-Verl., Heidelberg

Gaul W, Vincent D (2017) Evaluation of the evolution of relationships between topics over time. Advances in Data Analysis and Classification 11(1):159–178, DOI 10.1007/s11634-016-0241-2

Gayle HM, Blake BM (1980) Coal in west virginia: Geology and current mining trends. AAPG Bulletin 64(8):1287–1288

Gerken JM, Moehrle MG (2012) A new instrument for technology monitoring: Novelty in patents measured by semantic patent analysis. Scientometrics 91(3):645–670, DOI 10.1007/s11192-012-0635-7

Geum Y, Jeon J, Seol H (2013) Identifying technological opportunities using the novelty detection technique: A case of laser technology in semiconductor manufacturing. Technology Analysis & Strategic Management 25(1):1–22, DOI 10.1080/09537325.2012.748892

Glänzel W, Thijs B (2012) Using 'core documents' for detecting and labelling new emerging topics. Scientometrics 91(2):399–416, DOI 10.1007/s11192-011-0591-7

González-Alcaide G, Llorente P, Ramos JM (2016) Bibliometric indicators to identify emerging research fields: Publications on mass gatherings. Scientometrics 109(2):1283–1298, DOI 10.1007/s11192-016-2083-2

Goorha S, Ungar L (2010) Discovery of significant emerging trends. Proceedings of the 16th ACM International Conference on Knowledge Discovery and Data Mining pp 57–64, DOI 10.1145/1835804.1835815

von der Gracht HA, Vennemann CR, Darkow IL (2010) Corporate foresight and innovation management: A portfolio-approach in evaluating organizational development. Learning the Future Faster 42(4):380–393, DOI 10.1016/j.futures.2009.11.023

Grandjean N, Charpiot B, Pena CA, Peitsch MC (2005) Competitive intelligence and patent analysis in drug discovery: Mining the competitive knowledge bases and patents. Drug Discovery Today: Technologies 2(3):211–215, DOI 10.1016/j.ddtec.2005.08.007

Guo H, Weingart S, Börner K (2011) Mixed-indicators model for identifying emerging research areas. Scientometrics 89(1):421–435, DOI 10.1007/s11192-011-0433-7

Hennig P, Berger P, Meinel C (2013) Identify emergent trends based on the blogosphere. 2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT) 3, DOI 10.1109/WI-IAT.2013.147

Hiltunen E (2008) The future sign and its three dimensions. Futures 40(3):247–260, DOI 10.1016/j.futures.2007.08.021

Huang J, Peng M, Wang H, Cao J, Gao W, Zhang X (2017) A probabilistic method for emerging topic tracking in microblog stream. World Wide Web 20(2):325–350, DOI 10.1007/s11280-016-0390-4

Huang MH, Chang CP (2014) Detecting research fronts in oled field using bibliographic coupling with sliding window. Scientometrics 98(3):1721–1744, DOI 10.1007/s11192-013-1126-1

Huang Y, Zhang Y, Ma J, Porter A, Wang X (2015) Tracing technology evolution pathways by combining tech mining and patent citation analysis. 2015 Portland International Conference on Management of Engineering & Technology

Imran H, Sharan A (2010) A framework for automatic query expansion. In: International Conference on Web Information Systems and Mining, Springer Berlin Heidelberg, pp 386–393, DOI 10.1007/978-3-642-16515-3_48

Jeong Y, Yoon B (2015) Development of patent roadmap based on technology roadmap by analyzing patterns of patent development. Technovation 39-40:37–52, DOI 10.1016/j.technovation.2014.03.001

Jun S, Park SS, Jang DS (2012a) Patent management for technology forecasting: A case study of the bio-industry. Journal of Intellectual Property Rights 17(6):539–546

Jun S, Sung Park S, Sik Jang D (2012b) Technology forecasting using matrix map and patent clustering. Industrial Management & Data Systems 112(5):786–807, DOI 10.1108/02635571211232352

Kämpf M, Tessenow E, Kenett DY, Kantelhardt JW (2015) The detection of emerging trends using wikipedia traffic data and context networks. PLOS ONE 10(12):e0141,892, DOI 10.1371/journal.pone.0141892

Keller J, von der Gracht HA (2014) The influence of information and communication technology (ict) on future foresight processes: Results from a delphi survey. Technological

Forecasting and Social Change 85(C):81–92, DOI 10.1016/j.techfore.2013.07.010
Kim B, Gazzola G, Lee JM, Kim D, Kim K, Jeong MK (2014) Inter-cluster connectivity analysis for technology opportunity discovery. Scientometrics 98(3):1811–1825, DOI 10.1007/s11192-013-1097-2
Kim D, Kim D, Hwang E, Rho S (2015a) Twittertrends: A spatio-temporal trend detection and related keywords recommendation scheme. Multimedia Systems 21(1):73–86, DOI 10.1007/s00530-013-0342-0
Kim J, Hwang M, Jeong DH, Jung H (2012) Technology trends analysis and forecasting application based on decision tree and statistical feature analysis. Expert Systems with Applications 39(16):12,618–12,625, DOI 10.1016/j.eswa.2012.05.021
Kim N, Lee H, Kim W, Lee H, Suh JH (2015b) Dynamic patterns of industry convergence: Evidence from a large amount of unstructured data. Research Policy 44(9):1734–1748, DOI 10.1016/j.respol.2015.02.001
Kim S, Kim YE, Bae KJ, Choi SB, Park JK, Koo YD, Park YW, Choi HK, Kang HM, Hong SW (2013) Nest: A quantitative model for detecting emerging trends using a global monitoring expert network and bayesian network. Futures 52:59–73, DOI 10.1016/j.futures.2013.08.004
Kitchenham B, Charters S (2007) Guidelines for performing systematic literature reviews in software engineering: Technical report ebse 2007-001. keele university and durham university joint report
Kontostathis A, Galitsky LM, Pottenger WM, Roy S, Phelps DJ (2004) A survey of emerging trend detection in textual data mining DOI 10.1007/978-1-4757-4305-0_9
Kuosa T (2010) Futures signals sense-making framework (fssf): A start-up tool to analyse and categorise weak signals, wild cards, drivers, trends and other types of information. Futures 42(1):42–48, DOI 10.1016/j.futures.2009.08.003
Lau JY (2012) On-line trend analysis with topic models. Proceedings of the 24th International Conference on Computational Linguistics (24)
Lee C, Jeon J, Park Y (2011) Monitoring trends of technological changes based on the dynamic patent lattice: A modified formal concept analysis approach. Technological Forecasting and Social Change 78(4):690–702, DOI 10.1016/j.techfore.2010.11.010
Lee JY, Kim H, Kim PJ (2010) Domain analysis with text mining: Analysis of digital library research trends using profiling methods. Journal of Information Science 36(2):144–161, DOI 10.1177/0165551509353251
Lee S, Yoon B, Park Y (2009) An approach to discovering new technology opportunities: Keyword-based patent map approach. Technovation 29(6–7):481–497, DOI 10.1016/j.technovation.2008.10.006
Lee WH (2008) How to identify emerging research fields using scientometrics: An example in the field of information security. Scientometrics 76(3):503–525, DOI 10.1007/s11192-007-1898-2
Lee Y, Kim SY, Song I, Park Y, Shin J (2014) Technology opportunity identification customized to the technological capability of smes through two-stage patent analysis. Scientometrics 100(1):227–244, DOI 10.1007/s11192-013-1216-0
Lent B, Agrawal R, Srikant R (1997) Discovering trends in text databases. Proceedings of the Third International Conference on Knowledge Discovery and Data Mining pp 227–230
Liu DR, Shih MJ, Liau CJ, Lai CH (2009) Mining the change of event trends for decision support in environmental scanning. Expert Systems with Applications 36(2):972–984, DOI 10.1016/j.eswa.2007.10.016
Lu Y, Zhang P, Liu J, Li J, Deng S (2013) Health-related hot topic detection in online communities using text clustering. PLOS ONE 8(2):e56,221, DOI 10.1371/journal.pone.

0056221
Luo J, Pan X, Zhu X (2015) Identifying digital traces for business marketing through topic probabilistic model. Technology Analysis & Strategic Management 27(10):1176–1192, DOI 10.1080/09537325.2015.1061118
Ma J, Porter AL (2015) Analyzing patent topical information to identify technology pathways and potential opportunities. Scientometrics 102(1):811–827, DOI 10.1007/s11192-014-1392-6
Madani F (2015) 'technology mining' bibliometrics analysis: Applying network analysis and cluster analysis. Scientometrics 105(1):323–335, DOI 10.1007/s11192-015-1685-4
Mayer JH, Steinecke N, Quick R (2011) Improving the applicability of environmental scanning systems: State of the art and future research. In: Nüttgens M (ed) Governance and sustainability in information systems, IFIP Advances in Information and Communication Technology, vol 366, Springer, Berlin and London, pp 207–223, DOI 10.1007/978-3-642-24148-2_13
Mei Q, Zhai C (2005) Discovering evolutionary theme patterns from text: An exploration of temporal text mining. In: Grossman R, Bayardo R, Bennett K (eds) Proceeding of the eleventh ACM SIGKDD International Conference, pp 198–207, DOI 10.1145/1081870.1081895
Milanez DH, de Faria LIL, do Amaral RM, Leiva DR, Gregolin JAR (2014) Patents in nanotechnology: An analysis using macro-indicators and forecasting curves. Scientometrics 101(2):1097–1112, DOI 10.1007/s11192-014-1244-4
Moreira ALM, Hayashi TWN, Coelho GP, da Silva, Ana Estela Antunes (2015) A clustering method for weak signals to support anticipative intelligence. International Journal of Artificial Intelligence and Expert Systems (IJAE) 6(1)
Mryglod O, Holovatch Y, Kenna R, Berche B (2016) Quantifying the evolution of a scientific topic: Reaction of the academic community to the chornobyl disaster. Scientometrics 106(3):1151–1166, DOI 10.1007/s11192-015-1820-2
Murtaza SS, Khreich W, Hamou-Lhadj A, Bener AB (2016) Mining trends and patterns of software vulnerabilities. Journal of Systems and Software 117:218–228, DOI 10.1016/j.jss.2016.02.048
Nasraoui O, Rojas C, Cardona C (2006) A framework for mining evolving trends in web data streams using dynamic learning and retrospective validation. Computer Networks 50(10):1488–1512, DOI 10.1016/j.comnet.2005.10.021
Nguyen KL, Byung-Joo Shin, Seong Joon Yoo (2016) Hot topic detection and technology trend tracking for patents utilizing term frequency and proportional document frequency and semantic information. In: 2016 International Conference on Big Data and Smart Computing (BigComp), pp 223–230, DOI 10.1109/BIGCOMP.2016.7425917
Noh H, Song YK, Lee S (2016) Identifying emerging core technologies for the future: Case study of patents published by leading telecommunication organizations. Telecommunications Policy 40(10-11):956–970, DOI 10.1016/j.telpol.2016.04.003
Nohuddin PNE, Sunayama W, Christley R, Coenen F, Setzkorn C (2014) Trend mining in social networks: From trend identification to visualization. Expert Systems 31(5):457–468, DOI 10.1111/exsy.12024
Palomino MA, Vincenti A, Owen R (2013) Optimising web–based information retrieval methods for horizon scanning. Foresight 15(3):159–176, DOI 10.1108/fs-10-2011-0045
Park H, Kim E, Bae KJ, Hahn H, Sung TE, Kwon HC (2011) Detection and analysis of trend topics for global scientific literature using feature selection based on gini-index. 2011 IEEE International Conference on Tools with Artificial Intelligence pp 965–969, DOI 10.1109/ICTAI.2011.166

Park S, Kim J, Lee H, Jang D, Jun S (2016) Methodology of technological evolution for three-dimensional printing. Industrial Management & Data Systems 116(1):122–146, DOI 10.1108/IMDS-05-2015-0206

Parker J, Wei Y, Yates A, Frieder O, Goharian N (2013) A framework for detecting public health trends with twitter. In: Rokne J, Faloutsos C (eds) Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2013), pp 556–563, DOI 10.1145/2492517.2492544

Pinto JCL, Chahed T, Altman E (2015) Trend detection in social networks using hawkes processes. In: Pei J, Silvestri F, Tang J (eds) The 2015 IEEE/ACM International Conference, pp 1441–1448, DOI 10.1145/2808797.2814178

Porter AL, Youtie J, Shapira P, Schoeneck DJ (2008) Refining search terms for nanotechnology. Journal of Nanoparticle Research 10(5):715–728, DOI 10.1007/s11051-007-9266-y

Preschitschek N, Niemann H, Lenker J, Moehrle MG (2013) Anticipating industry convergence: semantic analyses vs ipc co-classification analyses of patents. Foresight 15(6):446–464, DOI 10.1108/FS-10-2012-0075

Rill S, Reinel D, Scheidt J, Zicari RV (2014) Politwi: Early detection of emerging political topics on twitter and the impact on concept-level sentiment analysis. Knowledge-Based Systems 69:24–33, DOI 10.1016/j.knosys.2014.05.008

Robertson S (2004) Understanding inverse document frequency: on theoretical arguments for idf. Journal of Documentation 60(5):503–520, DOI 10.1108/00220410410560582

Rodriguez A, Tosyali A, Kim B, Choi J, Lee JM, Coh BY, Jeong MK (2016) Patent clustering and outlier ranking methodologies for attributed patent citation networks for technology opportunity discovery. IEEE Transactions on Engineering Management 63(4):426–437, DOI 10.1109/TEM.2016.2580619

Rohrbeck R, Bade M (2012) Environmental scanning, futures research, strategic foresight and organizational future orientation: A review, integration, and future research directions: Ispim annual conference, barcelona, spain

Rohrbeck R, Thom N, Arnold HM (2015) It tools for foresight: The integrated insight and response system of deutsche telekom innovation laboratories. Technological Forecasting and Social Change 97(8):115–126

Santo Md, Coelho GM, dos Santos DM, Filho LF (2006) Text mining as a valuable tool in foresight exercises: A study on nanotechnology: A study on nanotechnology. Technological Forecasting and Social Change 73(8):1013–1027, DOI 10.1016/j.techfore.2006.05.020

Saritas O, Smith JE (2011) The big picture – trends, drivers, wild cards, discontinuities and weak signals. Futures 43(3):292–312, DOI 10.1016/j.futures.2010.11.007

Schult R, Spiliopoulou M (2006) Discovering emerging topics in unlabelled text collections. In: Manolopoulos Y (ed) Advances in databases and information systems, Lecture Notes in Computer Science, vol 4152, Springer, Berlin [u.a.], pp 353–366, DOI 10.1007/11827252_27

Schwarz JO (2005) Pitfalls in implementing a strategic early warning system. Foresight 7(4):22–30, DOI 10.1108/14636680510611813

Shibata N, Kajikawa Y, Takeda Y, Sakata I, Matsushima K (2011) Detecting emerging research fronts in regenerative medicine by the citation network analysis of scientific publications. Technological Forecasting and Social Change 78(2):274–282, DOI 10.1016/j.techfore.2010.07.006

Shih MJ, Liu DR, Hsu ML (2010) Discovering competitive intelligence by mining changes in patent trends. Expert Systems with Applications 37(4):2882–2890, DOI 10.1016/j.eswa.2009.09.001

Song M, Kim MC, Jeong YK (2014) Analyzing the political landscape of 2012 korean presidential election in twitter. IEEE Intelligent Systems 29(2):18–26, DOI 10.1109/MIS.2014.20

Steinecke NC, Quick R, Mohr T (2011) Environmental scanning systems: State of the art and first instantiation. PACIS 2011 Proceedings

Takahashi T, Tomioka R, Yamanishi K (2014) Discovering emerging topics in social streams via link-anomaly detection. IEEE Transactions on Knowledge and Data Engineering 26(1):120–130, DOI 10.1109/TKDE.2012.239

Tho QT, Hui SC, Fong A (2003) Web mining for identifying research trends. In: International Conference on Asian Digital Libraries, pp 290–301, DOI 10.1007/978-3-540-24594-0_28

Thorleuchter D, van den Poel D (2013) Weak signal identification with semantic web mining. Expert Systems with Applications 40(12):4978–4985, DOI 10.1016/j.eswa.2013.03.002

Thorleuchter D, Scheja T, van den Poel D (2014) Semantic weak signal tracing. Expert Systems with Applications 41(11):5009–5016, DOI 10.1016/j.eswa.2014.02.046

Trappey CV, Wu HY, Taghaboni-Dutta F, Trappey AJ (2011) Using patent data for technology forecasting: China rfid patent analysis. Advanced Engineering Informatics 25(1):53–64, DOI 10.1016/j.aei.2010.05.007

Tu YN, Hsu SL (2016) Constructing conceptual trajectory maps to trace the development of research fields. Journal of the Association for Information Science and Technology 67(8):2016–2031, DOI 10.1002/asi.23522

Tu YN, Seng JL (2012) Indices of novelty for emerging topic detection. Information Processing & Management 48(2):303–325, DOI 10.1016/j.ipm.2011.07.006

Veugelers M, Bury J, Viaene S (2010) Linking technology intelligence to open innovation. Technological Forecasting and Social Change 77(2):335–343, DOI 10.1016/j.techfore.2009.09.003

Vidhya KA, Aghila G (2010) Text mining process, techniques and tools: an overview. International Journal of Information Technology and Knowledge Management (2):613–622

Wang J, Li L, Tan F, Zhu Y, Feng W (2015a) Detecting hotspot information using multi-attribute based topic model. PLOS ONE 10(10):e0140,539, DOI 10.1371/journal.pone.0140539

Wang MY, Chang DS, Kao CH (2010) Identifying technology trends for r&d planning using triz and text mining. R&D Management 40(5):491–509, DOI 10.1111/j.1467-9310.2010.00612.x

Wang X, McCallum A (2006) Topics over time: A non-markov continuous-time model of topical trends. 2006 Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining DOI 10.1145/1150402.1150450

Wang X, Cheng Q, Lu W (2014) Analyzing evolution of research topics with neviewer: A new method based on dynamic co-word networks. Scientometrics 101(2):1253–1271, DOI 10.1007/s11192-014-1347-y

Wang X, Qiu P, Zhu D, Mitkova L, Lei M, Porter AL (2015b) Identification of technology development trends based on subject-action-object analysis: The case of dye-sensitized solar cells. Technological Forecasting and Social Change 98:24–46, DOI 10.1016/j.techfore.2015.05.014

Wanner F, Stoffel A, Jäckle D, Kwon BC, Weiler A, Keim DA (2014) State-of-the-Art Report of Visual Analysis for Event Detection in Text Data Streams. The Eurographics Association

Weenen TC, Ramezanpour B, Pronker ES, Commandeur H, Claassen E (2013) Food-pharma convergence in medical nutrition- best of both worlds? PLOS ONE 8(12):e82,609, DOI 10.1371/journal.pone.0082609

Wetzker R, Zimmermann C, Bauckhage C (2010) Detecting trends in social bookmarking systems. International Journal of Data Warehousing and Mining 6(1):38–57, DOI 10.4018/jdwm.2010090803

Woon WL, Madnick S (2012) Semantic distances for technology landscape visualization. Journal of Intelligent Information Systems 39(1):29–58, DOI 10.1007/s10844-011-0182-3

Woon WL, Henschel A, Madnick S (2009) A framework for technology forecasting and visualization. International Conference on Innovations in Information Technology (IIT) pp 115–159, DOI 10.1109/IIT.2009.5413768

Wu FS, Shiu CC, Lee PC, Su HN (2010) Integrated methodologies for mapping and forecasting science and technology trends: A case of etching technology. 2010 Technology Management for Global Economic Growth (PICMET) pp 1–23

Xie W, Zhu F, Jiang J, Lim EP, Wang K (2016) Topicsketch: Real-time bursty topic detection from twitter. IEEE Transactions on Knowledge and Data Engineering 28(8):2216–2229, DOI 10.1109/TKDE.2016.2556661

Yang L, Lin H, Lin Y, Liu S (2016) Detection and extraction of hot topics on chinese microblogs. Cognitive Computation 8(4):577–586, DOI 10.1007/s12559-015-9380-6

Yoon B, Park Y (2007) Development of new technology forecasting algorithm: Hybrid approach for morphology analysis and conjoint analysis of patent information. IEEE Transactions on Engineering Management 54(3):588–599, DOI 10.1109/TEM.2007.900796

Yoon J, Kim K (2012) Detecting signals of new technological opportunities using semantic patent analysis and outlier detection. Scientometrics 90(2):445–461, DOI 10.1007/s11192-011-0543-2