UNIVERSITY OF TWENTE.

Faculty of Behavioral Management
and Social Sciences

Automatic Weak Signal Detection and Forecasting

Tim Gutsche

Master Thesis

M.Sc. Business Administration

M.Sc. Innovation Management & Entrepreneurship

Date of submission: 24.08.2018

University of Twente:

First supervisor: Dr. Fons Wijnhoven

Second supervisor: Dr. Matthias de Visser

Technische Universität Berlin:

Supervisor: Birgit Peña Häufler

Abstract

Analyzing foresight approaches combined with weak signal mining reveals three significant shortcomings. First, foresight approaches are built on snapshot data, qualitative methods and lack of automation. All of which is insufficient in a constantly changing environment. Second, previous scholars, which combine foresight with weak signals focus on the sole detection of weak signals. Third, neither detecting weak signals, nor waiting for them to become strong signals is sufficient for strategic decision making. This study addresses these shortcomings with and aims to support strategic issue management by automatically detecting weak signals and forecasting their appearance through temporal web mining and time-series analysis. The usefulness of the study and its results are exemplified by a case study analyzing the environment of web conferencing solutions. The results show that the appearance of weak signals can be forecasted with a high f1-score and practitioners value the resulting system in terms of usability and utility.

Outline of the Thesis

List of Figures

List of Abbreviations

RTF: Relevant term frequency

ARTF: Average relevant term frequency

ROC: Receiver operating characteristics

AUC: Area under the curve

LDA: Latent Dirichlet Allocation

1 Introduction

Traditionally, strategic planning is the process of "converting environmental information about strategic discontinuities into concrete action plans, programs and budgets" (Ansoff, 1975, p. 22). However, constant change in the environment with incomplete and asynchronous information poses threats and opportunities to businesses and their strategy activities. Therefore, companies are seeking strategic orientation to exploit growth opportunities while maintaining competitive advantage (Mühlroth & Grottke, 2018). In order to incorporate a systematic approach of thinking about the future and enrich the context for strategic planning, companies apply foresight (Voros, 2003). Foresight aims to support decision making by analyzing future trends, and technologies in the environment. It incorporates techniques such as brainstorming, trend extrapolation, expert panels and scenario analyses (Voros, 2003).

However, foresight methods have three crucial shortcomings. First, these approaches are usually built on qualitative approaches and expert opinions (Holopainen & Toivonen, 2012; Kayser & Blind, 2017). Second, and building upon the first shortcoming, methods based on expert opinions or qualitative work, can by nature not be conducted automatically (Keller & von der Gracht, 2014; Mühlroth & Grottke, 2018). Existing approaches rely on manual input and expert opinions in various steps of the process (Mühlroth & Grottke, 2018). Third, whenever companies decide to practice foresight, only the currently available information can be taken into consideration. Therefore, the methods are built on static information and are not adaptive to the changing environment with changing information (Mühlroth & Grottke, 2018). Whereas "companies need to navigate and act in an environment that is subject to constant change" (Mühlroth & Grottke, 2018, p. 644) the manual and expert-based process of foresight with these shortcomings is not sufficient to support companies with strategic decision making. Given these shortcomings, the identification of promising long-term business opportunities remains the primary challenge of strategic planning (Yoon, 2012).

One essential ingredient and the first step of the foresight process is the availability and collection of expertise and information sources (Voros, 2003). Therefore, over the last years, external data sources have been incorporated to complement foresight techniques. Sources like patent data and publication data are used to map the technology landscape (Aharonson & Schilling, 2016), create patent intelligence systems (Park, Kim, Choi, & Yoon, 2013), and discover emerging technologies (J. Kim, Hwang, Jeong, & Jung, 2012) or clusters (Breitzman & Thomas, 2015) to name a few. Apart from these data sources, recent applications of foresight

methods make little use of other data sources like web content (Kayser & Blind, 2017). Web content provides textual data which could be used to perceive the ongoing changes in the environment and make statements about the future (Kayser & Blind, 2017). Recent studies used those data sources to extract new, useful and interdisciplinary ideas (Thorleuchter & Van den Poel, 2016; Thorleuchter & Van Den Poel, 2013b; Thorleuchter, Van Den Poel, & Prinzie, 2010). In order to make web data useful for foresight, scholars suggest text mining. Advancements in processing power and natural language processing support the derivation of insights from text data (Hirschberg & Manning, 2015). As one of the scholars applying web data and text mining in order to conduct foresight, Yoon (2012) focuses on the detection of early indicators in datasets, so-called weak signals, to derive future business opportunities.

The detection of weak signals is a means of conducting foresight (Holopainen & Toivonen, 2012). Weak signals are "warnings (..), events and developments that are still too incomplete to permit an accurate estimation of their impact and/or to determine their full-fledged responses" (Ansoff, 1982, p.12; Hiltunen, 2008b, p. 248). Moreover, even though weak signals can become strong signals, which permit an accurate estimation of their impact, not every weak signal passes this transition. This situation between weak and strong signals poses a strategic paradox. Strategic actors can either wait for weak signals to become strong, which is suitable for planning, which is also likely too late for strategic decision making in corporate foresight (Mühlroth & Grottke, 2018). Alternatively, actors accept the incompleteness of weak signals, which is insufficient for strategic planning (Ansoff, 1975). On the basis of this paradox, foresight methods based on the sole detection of weak signals, are still uncertain. To the knowledge of the author previous scholars focused solely on the detection of weak signals. According to Ansoff (1975), this paradox can be resolved by strategic issue management. Strategic issue management incorporates the planning of a graduated response along the evolution of weak signals, implying a continuous monitoring process of weak signals.

To synthesize: First, previous foresight approaches are built on static information, qualitative methods and lack of automation (Mühlroth & Grottke, 2018). All of which is insufficient in a constantly changing environment. Second, weak signals and their likely transition to strong signals are uncertain. Neither detecting weak signals, nor waiting for them to become strong signals is sufficient for strategic decision making. Third, previous scholars, which combine foresight with weak signals focus solely on the detection of weak signals.

Based on this rationale, there is the need for continuous and automated weak signal monitoring combined with forecasting. To fill this research gap, the proposed work aims to

support strategic issue management by automatically detecting weak signals and forecasting their appearance through temporal web mining and time-series analysis.

This study wants to find out whether the occurrence of weak signals can be predicted, hence state which weak signals might be more important and relevant in the future. For this reason, the following research questions are derived: *How accurate can the occurrence of weak signals be predicted?* The leading sub-question for this research is, *what predictors affect the occurrence of weak signals?*

By answering the research question, the author contributes additional information and knowledge towards the occurrence of weak signals. Moreover, this study bears a new approach towards forecasting weak signals, whereas previous studies focused on the sole detection of such. Therefore, this research also advances technological foresight with an example of using temporal web data for forecasting. The work aims at improving strategic decision making by delivering an automated process of environmental scanning for managers without the need to process large amounts of data on the web manually. The aim is to give strategists an automatic tool to monitor and understand the environment and its dynamics with emerging and submerging signals.

The novelty of this research is the application of topic models, in conjunction with temporal web mining to identify weak signals and track their life cycle over time. Related research was already conducted by Thorleuchter and Van Den Poel (2013a) and Yoon (2012). However, while the former scholars did parts of the process manually, the latter used keyword-based mining techniques. Keyword-based applications are restrained to the fact, that keywords never change, the hereby applied topic models, are a better way of finding semantic clusters relating to specific topics even if they do not share a common word (Thorleuchter & Van Den Poel, 2013a).

2 Theory of weak signals

The theoretical framework for this study is based on the theory of weak signals, which explains the evolution of weak signals over time and finally the transformation into other phenomena like strong signals.

Initially, the theory of weak signals was introduced by Ansoff (1975). He distinguishes between strategic planning and strategic issue management. Strategic planning demands strong signals containing specific information that is available early enough. These requirements of strong signals regarding early and specific information are met, when trends develop steadily, implying that trend extrapolation is sufficient to forecast the future. This extrapolation, however, fails when the future development significantly departs from the past development, which is called strategic discontinuity. These discontinuities happen when firms are confronted with unfamiliar and threatening events. Those events which threaten profit or opportunities are understood as strategic surprises. Such surprises can be political changes, new technologies, economic phenomena or new competitors (Ansoff, 1975; Holopainen & Toivonen, 2012). According to Ansoff (1975) strategic issue management overcomes this shortcoming of strategic planning by planning a graduated response along the evolution of weak signals and being prepared for sudden discontinuities. Moreover, strategic issue management expands strategic planning by admitting weak signals and their incompleteness as a basis for decision making.

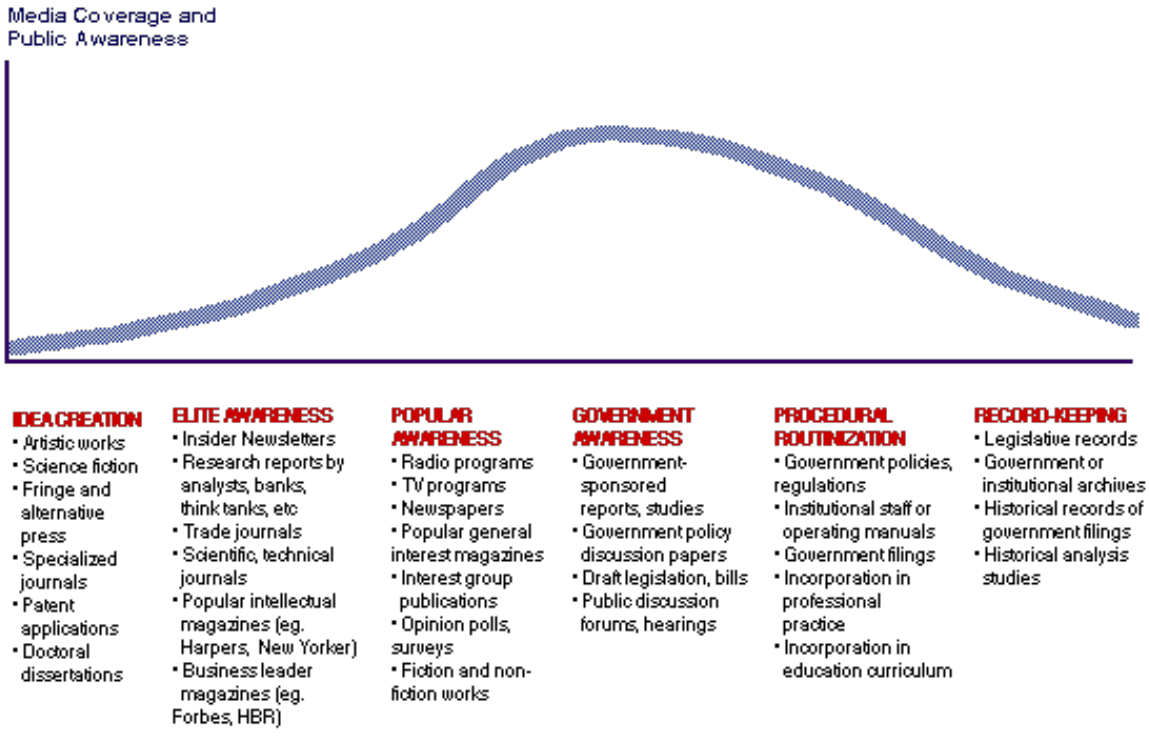

Figure 1: Information Lifecycle of emerging issues (Choo, 2002)

This study defines weak signals as "warnings (..), events and developments that are still too incomplete to permit an accurate estimation of their impact and/or to determine their complete responses" (Ansoff, 1982, p.12; Hiltunen, 2008b, p. 248). Throughout the lifecycle of weak signals, their information content improves, and their incompleteness decreases over time. The evolution of a weak signal to a strong signal, and further to a trend or even megatrend can be illustrated using the information lifecycle proposed by Choo (2002) (see Figure 1). In the course from a weak signal towards transforming into a strong signal, the weak signal may reach a wider audience and appears in several contexts as depicted in Figure 1 (Choo, 2002; Hiltunen, 2008a; Holopainen & Toivonen, 2012; Molitor, 2003). Apart from the sources illustrated in Figure 1, the Internet and especially blogs are potentially useful sources for finding weak signals (Hiltunen, 2007).

As suggested by definition, early in the life, the information is incomplete, and the future impact cannot be accurately estimated. Later the information becomes complete which allows better planning by the firm (Ansoff, 1975, 1980; Thorleuchter & Van Den Poel, 2013a).

While the information content of weak signals increases over time, once the information permits accurate estimation of the impact, a weak signal can be regarded as a strong signal. A strong signal differs from a weak signal by its greater probability of realization regarding the estimated impact (Holopainen & Toivonen, 2012). But not every signal might reach this state, weak signals which impacts are not existent later, are still regarded as weak signals. Even an important weak signal may not develop into an actual and impactful phenomenon because it is never perceived (Holopainen & Toivonen, 2012).

The evolution of weak signals into strong signals poses a strategic paradox. Strategic actors can either wait for weak signals to become strong, which is suitable for planning, which is also likely too late for strategic decision making in corporate foresight (Mühlroth & Grottke, 2018). Alternatively, actors accept the incompleteness of weak signals, which is not specific enough for strategic planning (Ansoff, 1975).

In order to operationalize the concept of weak signals and differentiate weak from strong signals Hiltunen (2008b) built upon the definition of Ansoff (1982) and created a framework which allows the determination of the weakness of a signal. The scholar calls this concept future signs, which is designed to function as a general model to understand the concept of weak signals. Future signs contain three dimensions with the following units:

- the *signal*: the number and/or visibility of signals
- the *issue*: the number of events
- the *interpretation*: the receiver's understanding of future sign's meaning

Hiltunen (2008b) argues that there are two objective dimensions of the sign consisting of the axes signal and issue. The objectivity, therefore, is derived from the fact that the number of events and signals are countable. The third and subjective dimension is the interpretation of the sign, which includes the context aspect. Interpretation is related to the receiver and interpreter of the sign. A sign strengthens when there is a rise in at least one of the dimensions.

Yoon (2012) applies the framework of Hiltunen (2008b) and evaluates the visibility of signals by measuring the occurrence frequencies of keywords. By deriving the occurrence frequency of keywords over a more extended period, Yoon (2012) can calculate the average occurrence frequency of a keyword and the increasing rate of the occurrence frequency. The analysis is built on two propositions: (1) Keywords of many occurrences in a collection are important and (2) recent appearances of keywords are more important than past appearances. The determination of weak signals follows the rationale that keywords with a low occurrence frequency but a high increasing rate can be classified as weak signals, respectively keywords exhibiting a high occurrence frequency and high increasing rate as strong signals. This means that both weak and strong signals exhibit growing visibility in terms of their occurrence frequencies. However, whereas the exposure of weak signals in terms of total occurrence is still low, strong signals already have a higher total occurrence frequency and are thus considered to be important and exposed to the external.

This study adopts the approach proposed by Yoon (2012) to detect weak signals and differentiate between weak and strong signals.

3 Methodological Foundations

3.1 Design Science

3.1.1 Definition

The research paradigm of this study is based on design science proposed by Hevner, March, Park, & Ram, (2004) and Peffers, Tuunanen, Rothenberger, & Chatterjee, (2007). Design science is a problem-solving paradigm which aims at creating and evaluating IT artifacts which solve an identified organizational problem (Hevner et al., 2004). The problem addressed in this study is primarily a business problem for managers and decision makers and is addressed using a technology-based solution. So, design science which aims to develop technology-based solutions to important business problems fits as a research paradigm for this study. The result of design science, an artifact, is not exempt from natural laws or behavioral theories. To the contrary, their creation relies on existing theories (Hevner et al., 2004). Choosing design science research in this study does not mean excluding behavioral or natural sciences.

While Hevner et al., (2004) created a set of guidelines and evaluation methods, Peffers et al. (2007) contributed a common framework to recognize and evaluate the results of such research. This framework is also applied here to guide the research.

3.1.2 Design Science Research Process

In order to conduct design science research, Peffers et al. (2007) have created a methodology framework for design science research in information systems. This framework describes the steps: problem identification and motivation, definition of objectives for a solution, design and development of artifacts, demonstration, evaluation and finally communication. This study follows this framework.

Problem identification and motivation: The problems with previous applications of weak signal mining identified in the previous chapters are four-fold. First, there is a lack of automation in the field. Most studies rely on manual inputs or expert opinions (Mühlroth & Grottke, 2018). Second, most studies do not incorporate web data, and especially not temporal web data (Kayser & Blind, 2017). Third, within weak signal mining techniques, there is a need for machine learning approaches which allow corpus updates. Corpus updates refer to the possibility of augmenting the underlying document collection with new documents without restarting the whole method from scratch. Approaches not allowing corpus updates are designed as one-time efforts rather than continuous detection and monitoring systems (Mühlroth & Grottke, 2018). Fourth, while weak signal identification is already done, to the knowledge of the author the forecasting of weak signals has not been approached.

Definition of objectives for a solution: The objective of this study is to develop a fully automated solution which works on a continuous basis with new data and is independent of manual interventions. This solution should inform decision makers about the current environment and give forecasts on weak signals. The primary challenge here lies in analyzing the information and deriving weak signals.

Design and Development of Artifacts: The artifacts in this study are built upon temporal web documents. Furthermore, to reach the objective, there are two different artifacts developed which help to interpret different levels of abstraction from the raw data. Practitioners can use the artifacts to learn about raw documents, coherent topics over a longer time horizon, and forecasts about important topics.

Demonstration: To demonstrate the design and the artifacts, this study applies the artifacts to the case of a company. The method and the artifacts are applied to the requirements and environment of the company.

Evaluation and Communication: The value delivered by the artifacts are evaluated with experts of the same company the artifacts are applied to. The underlying forecasting algorithm is evaluated using common performance metrics like precision, recall, and f1, which is a valid form of evaluating an algorithm for design science and weak signal identification (Hevner et al., 2004; Mühlroth & Grottke, 2018). Moreover, the implications for theory and practice are discussed as well.

3.2 Technology Foresight

3.2.1 Definition

In order to foresee the future, and shape innovation plans scholars, policymakers and managers use technology forecasting for over forty years (Miles, Meissner, Vonortas, & Carayannis, 2017). Foresight is about how to survive in an increasingly competitive environment (Voros, 2003). It includes a wide variety of activities such as anticipation, forecasting, strategic intelligence, futures research or technology roadmaps (Pietrobelli & Puppato, 2016). In order to fit technology forecasting in the process of strategy, Voros (2003) differentiates between strategic thinking, strategic planning, and strategic development. While strategic thinking is about exploring options, strategic planning incorporates analysis, breaking done a goal into steps and finally implementing actions. Strategic development takes care of making decisions and setting directions. In Voros (2003) perception, foresight is an element of strategic thinking, where it enriches and enhances the context within the strategy is developed, planned and executed. Therefore, technology foresight does not replace strategic planning.

3.2.2 Foresight process

In order to conduct technology foresight, Horton (1999) laid out a three-tier process which was later advanced by Voros (2003). This process which is also applied in this study consists of the steps: inputs, foresight work, and outputs.

Whereas the inputs step describes the gathering of information and scanning the environment, the foresight work is the central part of this process. It is comprised of analysis, interpretation, and prospection. The Analysis, therefore, is a preliminary stage with the goal of getting the variety of inputs in order. Interpretation thereafter aims at finding deeper structures and insights within the data. Prospection as the last part of the foresight work creates forward views. The outputs step which concludes the foresight process delivers tangible and intangible outputs. The primary purpose of this step is to deliver the insights to stimulate the thinking about options. Whereas tangible outputs are for instance reports or workshops, intangible outputs would include the changes in thinking.

This foresight process at this point has delivered an expanded perception of strategic options, which enriches and enhances the context where strategic planning can build upon (Voros, 2003).

3.3 Latent Dirichlet Allocation

Apart from the theory, the application of distinct methods is also part of this research. Which is reason enough to shortly introduce the core methods applied to deliver a deeper understanding of the research design.

Topic modeling is the research field around extracting semantic text clusters from a large corpus of textual, non-structured data. One of the first methods in this field was called latent semantic indexing (LSI) (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990). Later another method was introduced by Blei, Ng and Jordan (2003) called Latent Dirichlet Allocation (LDA). It is defined, that a topic is a distribution over a fixed vocabulary. Furthermore, it is assumed, that these topics are specified before any data has been generated. The statistical model reflects the intuition that documents exhibit multiple topics, and each document exhibits them in different proportion. Each word in each document is drawn from one of these topics, where the selected topic is chosen from the per-document distribution over topics. Hence all documents share the same set of topics, but each document exhibits different proportions of those topics (D. M. Blei, 2012). The proportion of a topic in a document is called topic probability. LDA is an unsupervised machine learning algorithm. Hence it is not dependent upon initial human input to identify topics. The only input necessary is the unstructured textual data and the number of expected topics.

To synthesize and clarify the use of technical terms in this research, the following terms are defined as follows:

- Topic Model: Topic modeling refers to the process of aligning a statistical model over a collection of documents to cluster them to a number of topics. A topic model is the result of this process. It contains a collection of topics.
- Topic: A topic is a semantic cluster of words, drawn from documents.
- Number of Topics: Is a necessary input for the topic model algorithm. It defined how many topics are contained in a topic model

Initially, topic modeling was applied to news articles in the context of information retrieval, other fields of applications are event detection or emerging trend detection (Pépin, Kuntz, Blanchard, Guillet, & Suignard, 2017). LDA or topic models in general, are also prominent data mining methods for detecting weak signals (Mühlroth & Grottke, 2018; Pépin et al., 2017; Thorleuchter & Van Den Poel, 2013a). The use of topic models considers the fact, that weak signals are formulated by different persons. In contrast to related research relying on keywords for weak signal detection (Yoon, 2012), topic models consider aspects of meaning rather than words. "It might be that two textual patterns representing weak signals are related to a specific topic even if they do not share a common word" (Thorleuchter & Van Den Poel, 2013a, p. 3). This study follows other scholars applying topic models for weak signal detection who implicitly assume, that every topic is a signal (Pépin et al., 2017; Thorleuchter & Van Den Poel, 2013a). The determination of the weakness of a signal follows the rationale outlined in section 2.

3.4 Supervised Machine Learning

Apart from using LDA as a method for extracting semantic clusters from a text, supervised machine learning algorithms for predicting the evolution of weak signals are also used. In order, to learn a set of rules from instances, machine learning requires a dataset, where instances are represented using the same set of features. If those features are given with known labels, inductive machine learning can learn a set of rules from these instances (Kotsiantis, 2007). Once a machine learning classifier is fitted to a dataset, which means it has learned the rules, the same classifier can be used to classify the label of new, unseen instances and apply the rules it has learned.

4 Design of a weak signal detection and prediction system

This section describes the design of an automated weak signal detection and forecasting system. First, a short overview of the whole design process with all its steps and alignment with the foresight process is given. The subsequent sections describe each of these steps in more detail.

4.1 Overview

As previously outlined, the design process follows the general foresight process with the steps inputs, foresight work, and outputs. The alignment of the foresight process with the design process is illustrated in Figure 2. In order for the system to work it needs data. In this study, this data is web content; hence the first step is data collection through temporal web mining. The collected data builds the foundation for the foresight work. In the first part, the analysis part, the data must be ordered, in this study, a typical text mining procedure is applied: After the tokenization, the terms are stemmed, stop words removed and bigrams extracted (Kayser & Blind, 2017). The most important part, the interpretation follows the analysis. The interpretation aims to find deeper structure in the data. This structure is derived, through the identification of topics by applying the topic model algorithm, onto the time-based corpus, and the subsequent connection of topics over the complete time horizon, which is referred to as topic chains. In the final part of the foresight work, a binary classifier is trained to predict the future occurrence of topics. Therefore, the features and labels are extracted from the dataset in order to train the classifier. After the main foresight work is done, the findings are used to design two artifacts.

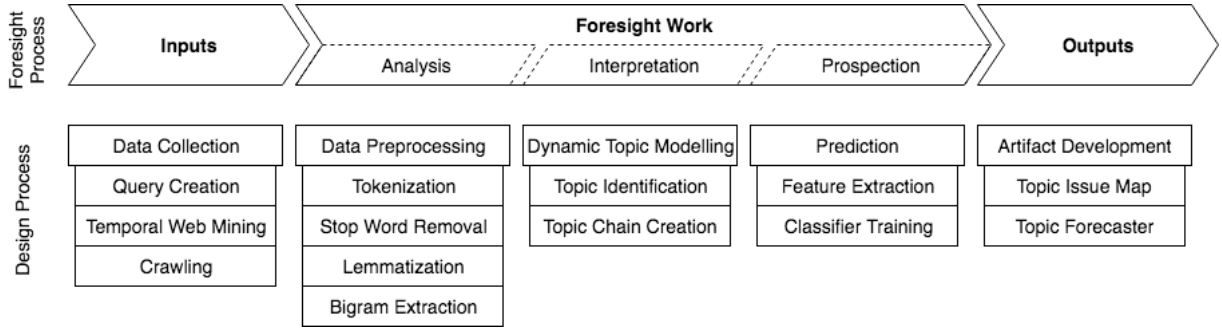

Figure 2: Alignment of foresight and design process

4.2 Data Collection

The collection of temporal web data requires three different inputs: a corpus of keywords, the time period as well as the duration of time slices within the time period. Together all three components form a set of queries. The result list of queries contains keywords together with the identifier of the time slice for which the keywords are searched. Therefore, the web

document search engine retrieves the most relevant results for the given keywords and the time slice of interest.

The processing of the search queries is done by a custom search engine from Google. The search engines results were further limited to include only articles, blogs, reviews and questions using the schema.org types. Schema.org is a community which aims at structuring the data on the internet. As a result, the service offers a vocabulary to indicate the type of data on the web (Khalili & Auer, 2013). Schema.org includes a variety of different Types. Apart from the information lifecycle presented in Figure 1 which illustrates newspapers as source of weak signals, Day and Schoemaker (2006) emphasize that the internet and especially blogs are good sources to find weak signals (Choo, 2002; Hiltunen, 2007). Therefore, the author chose to limit the results to only include articles, blogs, reviews, and questions, as these types are representing especially news articles and blogs.

The search results contain website metadata, as well as the hyperlink. In order, to retrieve the full text of these documents, a scraper is used, after that, the initial results are cleaned, and unfit samples are dismissed.

The data collection step concludes with a dataset of full-text dateable documents.

4.3 Data Preprocessing

In order to use the data for the proposed algorithms, it must be further processed. The preprocessing step aims to identify relevant terms within the corpus of full texts and prepare the data for the next task of topic identification. Therefore, the following steps are performed for every document in the dataset. This procedure includes tokenization, the removal of stop words, lemmatization and the extraction of bigrams and is common in various tasks of text learning.

First, the full text is tokenized, which refers to the separation of a long text into terms, whereas one term unit is defined as a word. After the terms are separated, numbers as well as short terms containing less than three characters, and stop words are removed using a list of common stop words for the English language. In the next step, the filtered list of terms is lemmatized, which reduces the word to the root form based on a dictionary using NLTK (Loper & Bird, 2002). Lastly, throughout the corpus bigrams which occur more often than ten times are identified, and the related terms are grouped as a new term. After this step, all the full-text web documents in the corpus are prepared for the next steps.

4.4 Dynamic Topic Modelling

4.4.1 Topic identification

For identifying topics within temporal data, scholars proposed dividing the temporal data into time windows of fixed duration (D. M. Blei & Lafferty, 2006; Sulo, Berger-Wolf, & Grossman, 2010). Therefore, the corpus containing the full date-stamped documents is separated into n disjoint time slices $\{T_i, \ldots, T_n\}$ of equal length. This step of processing the corpus in groups of time slices rather than processing the whole corpus is necessary because small and short-lived topics might vanish using the complete corpus. After separating the documents, for each time slice T_i, LDA is applied using the parameters based on the highest coherence measure c_v (discussed in detail in section 4.4.2) to all documents published in that time slice. This process yields a set of topic models $\{M_i, \ldots, M_n\}$ for each time slice T_i containing a set of Z_i of z time topics.

Training topic models for each time slice is not sufficient for detecting topics spanning over a longer time span than the length of a single time slice. Therefore, it is necessary to analyze the sequential evolution of topics over consecutive time slices. If topics evolve over time slices, the connection is referred to as a topic chain. In order to detect those topics, scholars proposed measuring the proximity of two topics $z \in Z_i$ and $z' \in Z_{i+1}$ by calculating the similarity $\delta(z, z')$ for both topic pairs (D. Kim & Oh, 2011; Pépin et al., 2017). High δ values correspond to strong relationships between topics over time. The choice of δ is discussed in detail in Section 4.4.3.

But, by attempting to find relationships only among the next consecutive time slice, and the length of those time slices is small, longer gaps are neglected, between the occurrence of topics. Therefore, not only the topics of two consecutive time slices are compared but of $\propto$ consecutive time slices. The factor $\propto$ describes the horizon, that topics can span without appearing, therefore from now on it is called spanning horizon. Taking the spanning horizon into account, the similarity of two topics $\delta(z, z')$ is measured with $z \in Z_i$ and $z' \in \bigcup_{j=1}^{\propto} Z_{i+j}$.

Measuring the similarity of topics produces a set of sequential connections between topics over time which is referred to as topic chains. The similarity, in this case, is an expression of the strength of the connection between two topics. For the final set of *topic chains,* it is also necessary to define a similarity threshold τ which indicates what similarity values are regarded as sufficient enough for connection between consecutive topics. Setting the threshold value too low leads to a large number of topics being similar, and hence connected to topic chains. A value too high leads to no connections at all. Subtracting the similarity value from one gives the dissimilarity which can be interpreted as the amount a topic can change over time. Figure 3

illustrates the process as levels of abstraction. Starting from raw documents, which are related to time slices, topics are derived and identified. Afterward, these topics are compared by their similarity and connected to topic chains.

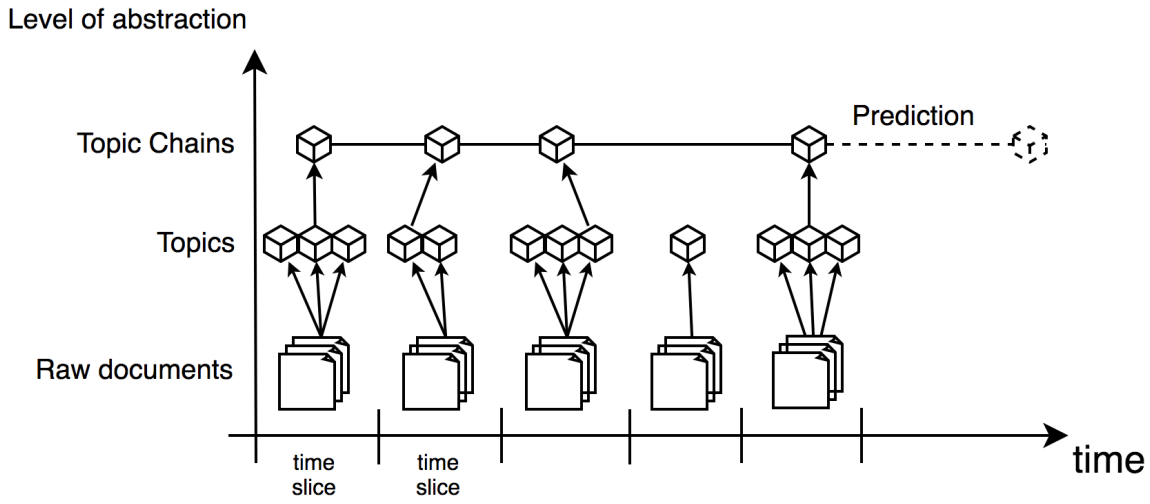

Figure 3: Overview of the method in levels of abstraction

4.4.2 Determination of number of topics

One challenge for finding optimal topic models is the number of topics k, which is a necessary input parameter for the topic model algorithm. If the number of k is too big, the algorithm will produce too many small, and similar topics. Choosing too few topics will produce very broad topics. To find the best number of topics, scholars propose evaluating the topic models by their topic coherence for different k (Chang, Gerrish, Wang, & Blei, 2009). This study follows this approach and apply the topic coherence measure c_v proposed by (Röder, Both, & Hinneburg, 2015).

To find the model with the highest coherence, an annealing approach is used. This approach implies the training of multiple models with different values for k and their subsequent evaluation by their coherence. The resulting model for k number of topics which yields the highest coherence is then chosen as the best model.

4.4.3 Selection of Distance measure

In order to build topic chains as described before, the calculation of the proximity of topic models and their topics in form of a similarity measure is necessary. Pépin, Kuntz, Blanchard, Guillet, and Suignard (2017) did related work, and they compared different measures for their application. The scholars compared different similarity measures by computing the edge proportion depending on the similarity threshold τ. Given all edges, for all the proximities between all topics calculated, the similarity threshold is a value which cuts off all the edge similarities which are less than this value. The result of this filtering over all edges

by the similarity threshold is a number of edges. The proportion of all edges to the edges over the threshold is then defined as the edge proportion (see Figure 4). From this depiction, the scholars chose the one similarity measure which is the most efficient filter while avoiding importation variations of the number of edges when τ varies only slightly. Their choice was the distance measure which presented a sigmoid based variation. This study follows this approach, calculates the edge proportion, and chooses the Hellinger distance, as the most suitable, because it also presents a sigmoid based variation.

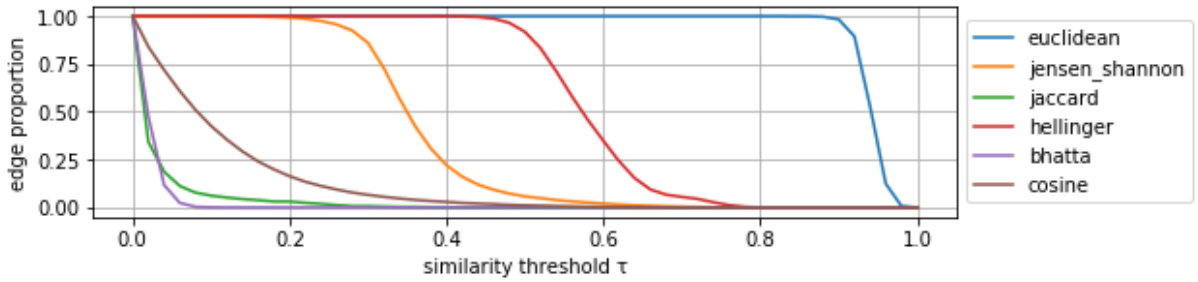

Figure 4: Number of connections in relation to similarity threshold τ.

4.5 Prediction

4.5.1 Classifier training

After the full set of topic chains is derived from the individual time-based topic models, and this set is filtered by a similarity threshold τ. The second step aims at learning what features predict whether a topic has a consecutive similar topic or not. This is a binary classification problem; therefore, each topic $z \in Z_i$ needs to be labeled according to the presence of a connection with a similarity value $\delta(z, z') > \tau$ to a consecutive topic $z' \in \bigcup_{j=1}^{\propto} Z_{i+j}$. Whereas a positive label indicates that a consecutive topic is present, which is labeled as *continued* a negative label indicates that the topic under inspection has no subsequent, similar topic, which is labeled *stopped.* This rule for extracting class labels neglects the number of connections to consecutive topics, whereas only the presence is evaluated but not the number of connections. Together with the labels, features for every topic are derived to form the binary classification dataset. A detailed description of the features follows in section 4.5.2.

To find the correct classifier for this problem, the automated machine learning tool TPOT is used, which uses genetic programming to find a best machine learning pipeline (Olson, Urbanowicz, et al., 2016; Olson, Bartley, Urbanowicz, & Moore, 2016). The resulting pipeline includes feature preprocessing, feature construction, model selection and parameter optimization. In order to eliminate human intervention in the design, this automated tool was chosen. Instead of a human expert determining the best machine learning classifier for the problem at hand, TPOT takes all classifiers into account to determine the best one, which is

chosen in the end. The pipeline is optimized using the f1-score. The f1-score was chosen because it is the harmonic mean of precision and recall and therefore balancing a score for relevancy and sensitivity of results (Sasaki, 2007).

4.5.2 Feature Selection

In order for a supervised machine learning classifier to learn a set of rules, it requires a dataset, where instances are represented using the same set of features. Therefore, first, it needs to be defined what features, could help in predicting the evolution of weak signals. The proposed approach assumes that signals are topics appearing in documents.

Rehashing the theory of weak signals especially the three-dimensions of signals proposed by Hiltunen (2008b), the following features can be derived for the prediction.

First, the scholar describes the issue as one of the dimensions describing the number of events of a signal. To quantify this idea, and derive features based on it, two features are defined:

1. The number of documents containing a certain topic
2. The topic frequency, which is the sum of topic probabilities divided by the number of documents.

The latter is also in line with Yoon (2012), who quantified the issue with the average document frequency which corresponds to the number of documents containing a keyword divided by the total number of documents.

Second, the other dimension in the framework is signal, describing the visibility of signals. Yoon (2012) quantified this dimension by the average term frequency. In this study, this dimension is translated into two features:

3. The total topic probability, which is the sum of all topic probabilities over all documents for a given topic.
4. The weighted length of documents, using the topic probability as weight and the number of terms in the document as length. This feature helps to estimate the number of terms related to a certain topic.

Third, Holopainen & Toivonen (2012) state that over the time, a weak signal starts to appear in different sources (see also Figure 1). This assumption is translated into one last feature:

5. The concentration of the topic distribution over documents. This feature is derived using the Herfindahl Hirschman Index of the share of topic probabilities, hence the sum of the squared share of a topic probability. A small value of this index is an

indicator that a topic is exhibited by a large number of different documents, whereas a higher value indicates the topic only appears in a few documents

4.6 Artifact Development

In his study, two objectives for the design of the artifacts were set. The artifacts should be able to inform decision makers first about the current environment and second make a forecast of weak signals in the future. In order to fulfill these two objectives, this study aims at designing two artifacts. Tableau[1] is used to build the artifacts. Tableau is a data visualization tool, capable of connecting to various data sources. Moreover, tableau enabled the interaction with the data and application of custom filters. This capability enriches the artifacts with interactivity and gives the practitioners more options in exploring the generated artifacts. Moreover, practitioners can apply their own set of assumptions in the form of parameters to the final artifacts.

4.6.1 Topic chain emergence map

The first artifact is based on Yoon's (2012) work of quantifying signals and their strength. As an artifact, the scholar developed a keyword emergence map. This study adopts this map for underlying topic models. Therefore, a topic chain emergence map is built which is depicted in Figure 5. Whereas, Yoon (2012) mapped the average term frequency and the time-weighted increasing rate of the average term frequency, this adaption calculates the sum of the term frequencies of the most relevant terms for all topics in a topic chain for the selected time slice and divide this sum by the number of documents in the time slice, which is referred to as average relevant term frequency (ARTF).

$$ARTF_{ij} = \frac{RTF_{ij}}{D_j}$$

RTF: frequency of relevant term of all topics in period j and topic chain i; D: total number of documents in period j

In order to calculate the relevance for terms of topics and derive the most relevant terms, this study follows Sievert and Shirley (2014). They propose the measure of relevancy as a superior method for topic interpretation over term probability. Within a user study, they showed that terms sorted by the relevance measure are more supportive of interpretation than terms sorted by term probability. Once the ARTF values are calculated for a topic chain and all periods, the increasing rate of these ARTF value for all periods as a geometric mean can be derived and the absolute average ARTF. Now, plotting both values on a graph results in the topic chain emergence map.

[1] https://www.tableau.com

According to Yoon (2012), such a map allows differentiating weak signals from strong signals. Therefore, weak signals and strong signals, both show a high increasing rate. Strong signals have a higher ARTF value. Hence they are more exposed compared to weak signals, showing lower values.

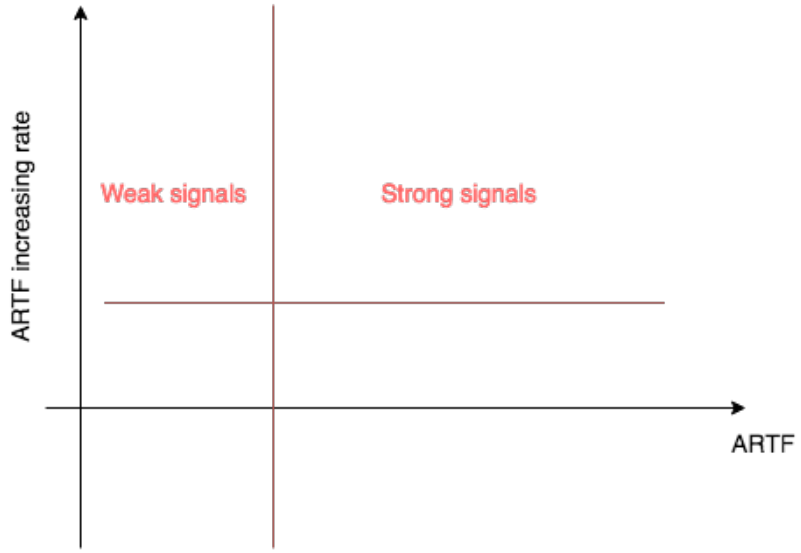

Figure 5: schematic topic chain emergence map

Additionally, the emergence map is enhanced with the insights gathered from the prediction to show those topic chains which are predicted to occur again. In order to facilitate interpreting the topics shown in the emergence map, the most relevant words with their frequency for each topic chain are also illustrated.

4.6.2 Topic Forecaster

The second artifact, which aims at giving decision makers a forecast of the future with predicted topics is primarily based on the termite system proposed by (Chuang, Manning, & Heer, 2012). Termite is a visual analysis system for human-centered topic modeling. It shows a topic term matrix combined with the most representative documents. This study adopts this artifact, to visualize for a chosen time slice those topics predicted to be present in the future with their most relevant words and most representative documents, which is referred to as topic forecaster. This artifact supports decision makers to focus on the most interesting and future bearing topics together with relevant documents.

5 Application Study / Demonstration

The case study applies the proposed methodology to support strategic decision makers at a company that offers web conferencing solutions for business clients. The companies' products and services are potentially affected by different market and technology trends. Moreover, a highly fragmented market with multiple competitors makes analyzing the market on a continuous basis difficult but important. The company seeks, therefore, an opportunity for automating this task and detect weak signals early. This case also builds the basis for a later evaluation with the same company.

5.1 Data collection and preprocessing

For this case, a long time period of 10 years is incorporated, hence the time period is set from January 2008 until April 2018. To have a larger number of topics and a higher probability of detecting weak signals, the time slice is chosen to be one month. Over the time period, this results in 124 time slices. The keywords were created together with industry experts from the partnering company and include the terms "web conferencing", "online meetings", "online presentation", "online collaboration services", "business streaming", "webinar", "webcast" and "web meeting". As the identifier for the search engine to filter by the date the parameters "metatags-date", "newsarticle-datepublished", "metatags-pubdate", "metatags-published_at" and "date" were used. The final set of queries contained 440 unique entries, which results from the combination of all keywords with all parameters for every year. These entries were entered into the custom search engine described in section 4.2.

Due to the decreasing relevance of search results on pages with higher page numbers, only the first two pages are incorporated (Thorleuchter & Van Den Poel, 2013b). The 440 queries resulted in 6,947 search results. These results are mainly composed of metadata and a link to a website. To filter the raw list of results, duplicate entries are removed which resulted in 5,936 entries, invalid date information within the metadata eliminated another 575 results resulting in 5,360 results. Then the crawling these results started. Obeying the website preferences for web crawlers (Kolay, 2008) and eliminating invalid links resulted in a number of 5,137 full-text web documents. As the last step, a shallow analysis of the texts of these full-text documents aimed at excluding documents with no or non-English text. As a result, 4,756 documents formed the final dataset, which is 68% of the initially retrieved results. Whereas only 147 samples had non-English text, the author decided against translating those and focus on English texts only. These steps result in the final collection of datable full-text documents for the given keywords, timeframe and duration of the time window.

In order to use the data for the proposed algorithms, it must be further processed. The preprocessing step aims to identify relevant terms within the corpus of full texts and prepare the data for the next task of topic modeling. Therefore, the following steps are performed for every document in the dataset. This procedure includes tokenization, the removal of stop words, lemmatization and the extraction of bigrams and is common in various tasks of text learning.

First, the full text is tokenized, which refers to the separation of a long text into terms, whereas one term unit is defined as a word. After the terms are separated, numbers as well as short terms containing less than three characters, and stop words are removed using a list of common stop words for the English language. In the next step, the filtered list of terms is lemmatized, which reduces the word to the root form based on a dictionary using NLTK (Loper & Bird, 2002). Lastly, throughout the corpus bigrams which occur more often than ten times are identified, and the related terms are grouped as a new term. After this step, all the full-text web documents in the corpus are prepared for the next steps.

5.2 Dynamic Topic Modelling

In order to find the best topic model for each time slice $\{T_i, \ldots, T_n\}$ an annealing approach was used, by training multiple models for a different number of topics k and evaluating their coherence measure c_v (Röder et al., 2015). Therefore, the process started by training models for every time slice with a predefined set for values of k resulting in a set of models $\{m_i, \ldots, m_n\}$ per time slice T_i. To have a broad range, every model was trained with values for k ranging from 5 to 85 in steps of 5. resulting in 16 models $\{m_1, \ldots, m_{16}\}$ per time slice. After this the annealing process worked as follows:

1. Find the best number of topics k where c_v_{max} of model m_i is max for each time slice T_i.
2. Find the distances Δ_{left} and Δ_{right} for the nearest models $m_{i-1}$and m_{i+1} in terms of their number of topics k.
3. Create new models $\{m'_i, \ldots, m'_n\}$ for k in $\frac{\Delta_{left}}{2}$, $\frac{\Delta_{right}}{2}$, $\frac{\Delta_{left}}{4}$ and $\frac{\Delta_{right}}{4}$ and search for new max value for c_v'.
4. If c_v' for any of the new models $\{m'_i, \ldots, m'_n\}$ is > c_v_{max}, restart at step 2 with the new c_v_{max}.
5. Return c_v_{max}.

This process applied to each time slice $\{T_i, \ldots, T_n\}$ results in a set of final topic models $\{M_i, \ldots, M_n\}$ with a total count of 3054 topics. Figure 6 shows the number of topics

for each of the final models illustrates the coherence measure c_v for these models and the topics.

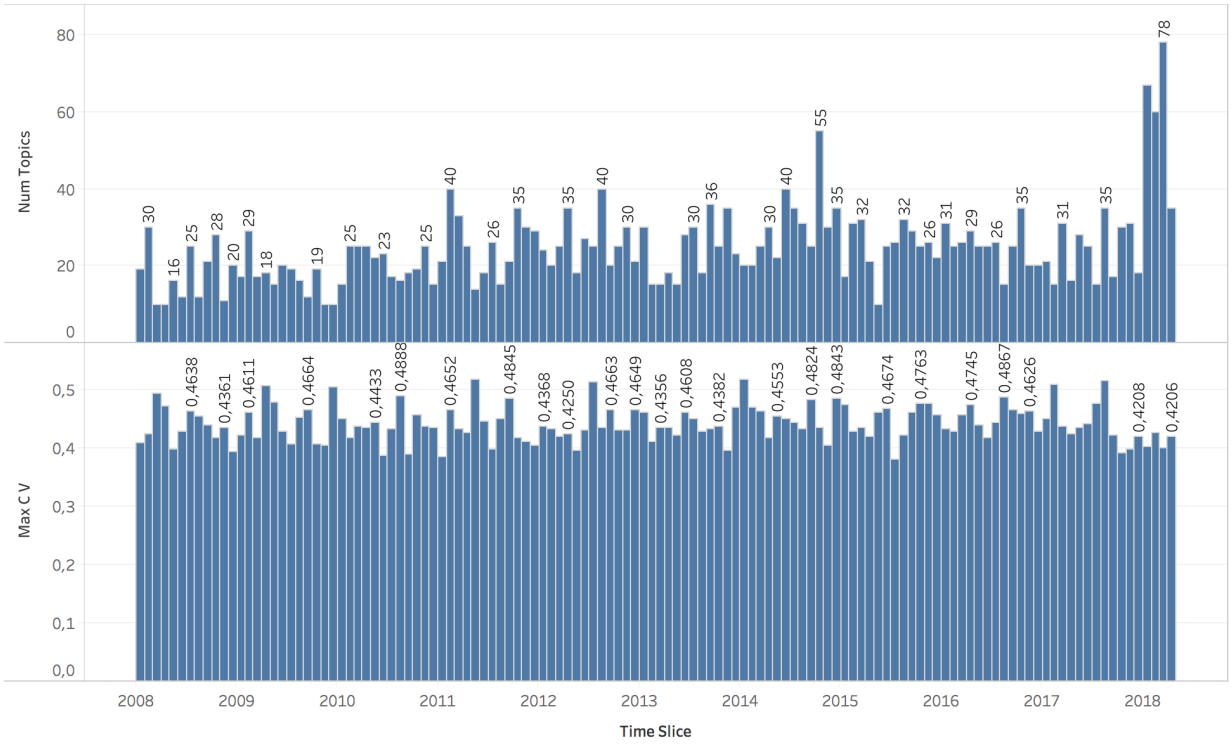

Figure 6: The number of topics (above) and maximum c_v value (below) of the optimal model per time slice

After the optimal models are evaluated, and all topics are retrieved, the next step aims at analyzing the sequential evolution of those topics over consecutive time slices to derive topic chains. Therefore, the similarity measure described in Section 4.4.3 is applied to all topics z and z' for the models M_i and $\bigcup_{j=1}^{\propto} M_{i+j}$ of all consecutive time slices within the spanning horizon T_i and $\bigcup_{j=1}^{\propto} T_{i+j}$. In accordance with similar studies which worked with time windows of months (Pépin et al., 2017), quarter (Greene & Cross, 2017) or half years (Yoon, 2012), the spanning horizon $\propto$ for this demonstration is set to 6 time slices, meaning that similar topics are searched within the next 6 time slices. Therefore, effectively the next half year is always included.

For the task of extracting features and labels, the author decided to take a similarity threshold of $\tau = 0.75$. The underlying assumption is that topics change but not drastically within the spanning horizon. Setting τ to low results in mixing different slightly different but distinct topics. Respectively, choosing a high value eliminates all connections at all.

In the final artifacts, the practitioners can set both values for $\propto$ and τ according to their assumptions. Therefore, a faster changing environment would translate to lower values for τ

and ∝. For an initial demonstration of the method, reasonable sample values for both parameters are used.

The final dataset includes features and labels for 3019 topics. Whereas the data includes 3,054 topics, for the topics of the last time slice a class label cannot be derived, because it is the last time slice, and no information about a consecutive time slice is known. Within the dataset, there are 2,539 topics with the class label *stopped*, which indicates that the topic has no similar topic in the subsequent time slice, and 480 positive class labels *continued* indicating the presence of at least one topic in the subsequent time slice.

5.3 Prediction Results

In order to answer the research questions, the classifier is evaluated using precision, recall, f1, and AUC. Moreover, to also answer the sub-question the importance of all features is analyzed to indicate which one contributes the most towards correct predictions.

To train the classifier and test it afterward, the dataset is separated into 33% test and 66% training data. Hence, 2,022 data points with 323 *continued* and 1699 *stopped* class labels form the training data, and 997 data points including 840 *stopped,* and 157 *continued* class labels form the testing data. The classifier is then trained using the training data and validated using the test data. TPOT is used to optimize a machine learning pipeline by evaluating the F1-score (Olson, Bartley, et al., 2016). Therefore, the TPOT classifier is trained for 100 generations and use five stratified folds as a cross-validation strategy. The optimal classifier is based on the gradient boosting classifier. The scores for this classifier for the test data can be found in Table 1.

Table 1: Classification Report for each label including Precision, recall, f1-score and support

Label	**Precision**	**Recall**	**F1-Score**	**Support**
stopped	0.97	0.96	0.96	840
continued	0.80	0.82	0.81	157
Average/Total	0.94	0.94	0.94	997

The recall score indicates that the classifier performs well on classifying the true positives from all positives. The precision score, however, shows that there are also some false positives, hence wrongly classified negative samples.

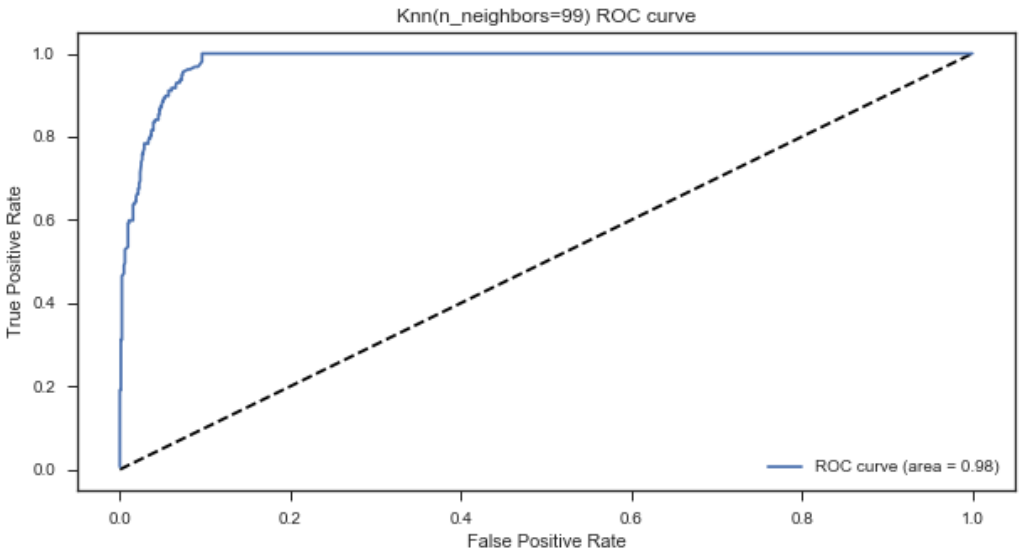

Figure 7: ROC curve for the classifier

Apart from the scores derived in Table 1, a receiver operating characteristics (ROC) graph is a popular tool for evaluating the performance of a classifier (see Figure 7). The curve depicts the tradeoff between the hit rates and the false alarm rates by plotting the false positive rate against the true positive rate (Fawcett, 2006). Calculating the area under the ROC curve (AUC) is a way to reduce the performance to a single scalar value, in this case, 0.98 which indicates the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance.

Additionally, to evaluate the overall performance of the classifier the importance of single feature contributions is also quantified. Therefore, a cross-validation with the already fitted classifier is performed, where only data from a single feature is used to predict the test data and derived the f1-score, precision, recall, and accuracy. The result of this cross-validation is illustrated in Table 2. and indicates how important a single feature is in predicting correct results.

Table 2: Cross-validation score using the trained classifier with a single feature

Feature	**F1-Score**	**Precision**	**Recall**	**Accuracy**
F1: number of documents showing the topic	0.78	0.67	0.94	0.92
F2: Topic Frequency	0.76	0.64	0.94	0.91
F3: Total Topic Probability	0.74	0.77	0.73	0.92
F4: Weighted Length of Documents	0.79	0.67	0.97	0.92
F5: Concentration of Topic Distribution over Documents	0.78	0.67	0.94	0.92

Taking a closer look into the results shows that the features *number of documents showing the topic* (F1), *the concentration of topic distribution over documents* (F5) and *weighted length of documents* (F4) are the most important feature for correct classifications. Those features produce the highest F1-Scores 0.78 and 0.79 respectively and when used alone. Apart from the most important feature, it is also clear that the features *total topic probability* (F3) and *topic frequency* (F2) are the weakest. However, every feature contributes to the classification. It can also be seen that the accuracy is always high, which can be attributed to the unbalanced class distribution, resulting in more negative samples than positive ones. The table also shows the effect of two of the features F3 and F4 on precision and recall. F3 seems to increase the precision, as it scores the highest precision when used alone. Whereas F4 has the highest effect on recall. These effects illustrate the importance of taking all features into account and not relying on a subset.

5.4 Artifact Derivation

After the data collection and preprocessing, the dynamic topic modeling and the prediction is done, the results of these steps can be transferred to build the proposed artifacts. Tableau is used to visualize the data gathered and generated through the processes and create the prototypes.

5.4.1 Topic chain emergence map

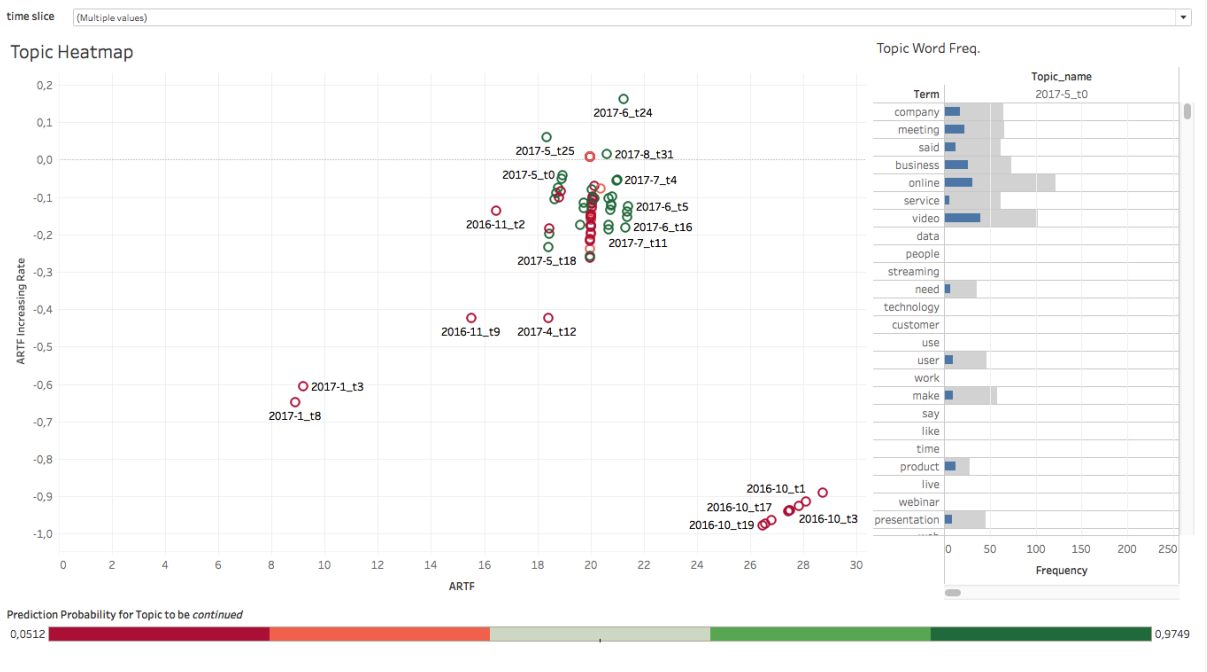

Figure 8: Topic chain emergence map artifact. It shows all topics (circles) of chosen time slices (top) and their prediction to occur in the future (left). The greener the circles are the higher the probability for the topics to appear in the future. The naming of the topics is derived by the time slice followed by a number identifying the topic number. Upon selecting a single circle, the top terms with their frequencies are shown (right). This view is generated using $\tau = 0.75$ *and* $\propto = 6$.

Blending the transitions and topic chains built in the dynamic topic modeling step together with the classifier able to predict the future occurrence for every topic, the topic chain emergence map can be built showing the growth and visibility of topics. Figure 8 shows the final artifact. From this view, it is clear that topics with a higher ARTF increasing rate are more likely to occur in the future. Whereas the circles only depict topics but not necessarily weak signals, the plot itself gives information about whether a topic is a weak or strong signal. Therefore, following Yoon (2012) topics with a higher increasing rate but a comparably low ARTF are considered weak signals. Respectively, more diffused topics, hence higher ARTF, and a high increasing rate are considered strong signals. Following this classification, it seems like the topics of the time slices selected here are rather strong than weak signals. Regardless of the actual data. This view enables the prediction of weak signals if there are any. Moreover, whereas this view depicts only one set of parameters, it is also possible for the practitioner to adapt the parameter settings to individual assumptions.

Upon selecting a single topic from the scatterplot on the left, the right view shows the most frequent terms for the selected topic (see Appendix A: Topic chain emergence map for a depiction). Whereas automatic labeling of the topics does not facilitate interpreting their content for a practitioner, the most frequent terms do.

5.4.2 Topic Forecaster

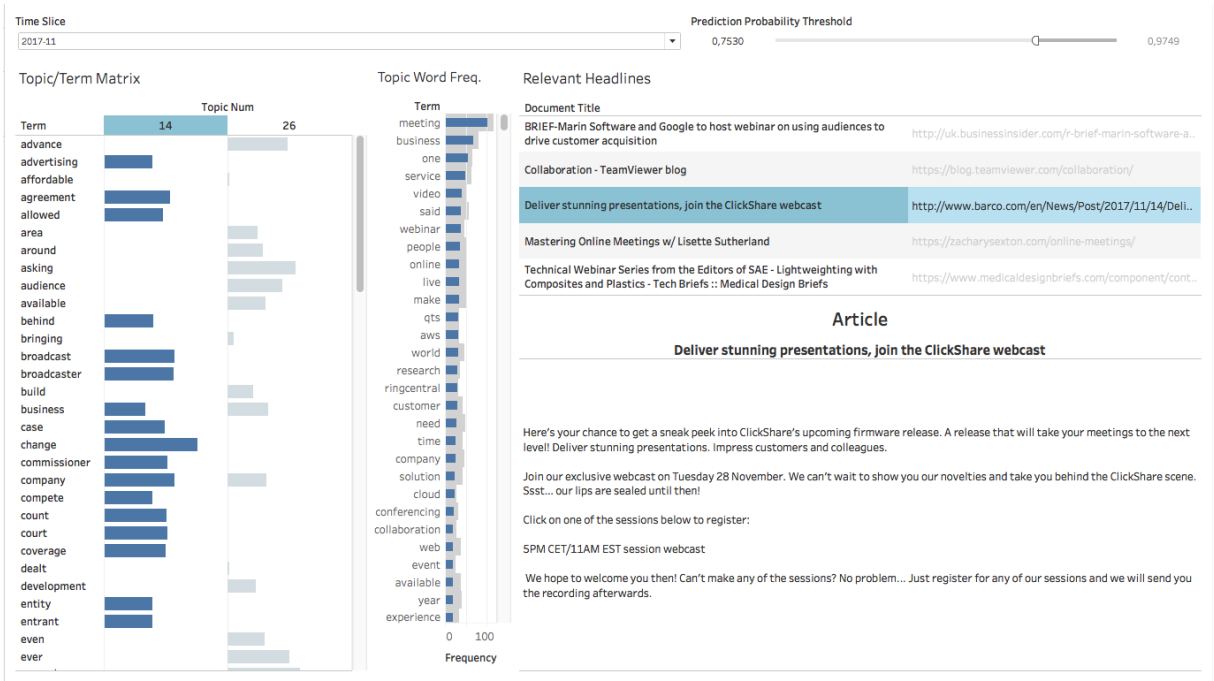

Figure 9: Topic forecaster artifact. By selecting a time slice (upper left), the user can see topics which are projected to occur in the future (left). The user can set the prediction probability threshold. In the middle, the most relevant terms with their term frequency is displayed. The right side shows relevant headlines and related articles. The user has also the possibility to directly click on the hyperlink to get to the webpage.

The topic chain emergence map is a tool for scanning the environment, exploring topics and discover weak or strong signals. However, it can also be distracting seeing topics, not relevant in the future. Therefore, another artifact is built, with the goal of giving decision makers focused attention on important and future relevant topics. In this topic forecaster (see Figure 9), the user is only confronted with topics that are predicted to occur again in the future. The user can control, what his threshold of prediction probability is (upper right). After selecting one of the predicted topics (left), the user can see relevant terms and their frequency (middle) and relevant headlines of documents showing the topic (right). By selecting one of the relevant headlines, the whole article is also displayed (lower right).

6 Discussion and Evaluation

6.1 Discussion and validation of artifacts

To verify the benefits of the proposed approaches and especially the artifacts, they have been evaluated with the firm delivering the case for the demonstration. The company operates in the B2B web conferencing and business streaming market. A market, with multiple competitors all offering similar solutions and low entry barriers opening the market for new entrants. The products and services of the company are potentially affected by different market and technology trends from within the market, but also other markets as well. Due to this situation, the company's challenge is to maintain existing market needs while exploring new niches and trends. To do so, the company seeks a low-cost method to automatically scan the environment and detect weak signals early and on a continuous basis.

The finalized artifacts developed in this study were shown to a senior executive with knowledge in the marketplace. The executive could try out the artifacts and interact with the results. Afterward, the utility, benefits, disadvantages and improvement ideas of the artifacts were discussed.

The most recognized benefits were time savings, level of automation, cost efficiency, objectivity, and level of insight. The fact that the artifacts are developed fully automatically on a continuous basis and without the need for a large infrastructure was most recognized. Moreover, the executive pointed out, that the level of abstraction from documents to topics and topic chains is useful, even if the concept is hard to get initially. It was highlighted, that the inclusion of all levels of abstraction, from raw documents, web links until topics helps in understanding the structure. This inclusion also facilitates trust in the system because it delivers transparency over the process. The depiction of topics as the most occurring terms also supports the interpretation of the topics. The topic chain emergence map specifically gives also objective measures and complements opinions and intuition of decision makers.

However, apart from the benefits, some disadvantages were also discussed. First, whereas the overview of the environment is very helpful, the decision maker expressed some initial doubts about the forecast. Second, as the system has parameters, they were not clear in the beginning. It was pointed out, that experience is necessary in order to fine-tuning the parameters. Third, the pure looking at the artifacts does not reveal any intelligence, the decision maker argued, that there is no automatic labeling of the topics. To derive insights, the user must interact with the artifact.

Lastly, the senior executive also expressed some ideas for improvements and further applications. For instance, the inclusion of more diverse data sources, or the application to other fields apart from environment scanning.

The findings discussed with the practitioner are also in line with Mühlroth & Grottke (2018). They also highlighted cost savings, objectivity and provision of a complementary overview as the main benefits outlined by experts evaluating weak signal mining techniques.

The objectives set for the design science process and aimed to satisfy with the artifacts were three-fold. First, the solution should be automated and work on a continuous basis. Second, it is meant to inform decision makers about the current environment, and third, give forecasts of weak signals for the future. The discussion with the practitioner shows that the artifacts are capable of satisfying all three objectives, which indicates according to Hevner et al. (2004) a complete and effective artifact.

6.2 Discussion and evaluation of the design

Practitioners in the field of strategic decision making can only evaluate the outcome of this method in forms of artifacts, but not the underlying method or algorithm. Therefore, we evaluated the prediction algorithm with known performance measures, which is a valid form of evaluation in design science (Hevner et al., 2004). The classifier used for the prediction had an average F1-Score of 0.94. Analyzing the ROC for this classifier shows that it is located in the upper-left corner with only a few false positive errors. This sort of classifier can be interpreted ‘conservative’ because it makes positive classifications only with strong evidence. Which is important in this case of a skewed class label distribution (Fawcett, 2006). Therefore, the resulting classifier is capable of determining whether a topic will be present in the next time slices. Which marks the general research question as answered.

Inspecting the predictive power of the single features revealed that three of the five features are especially powerful in predicting the outcome. Those three features are F1: number of documents showing the topic, F4: weighted length of documents and F5: concentration of topic distribution over documents. Whereas feature F1 and F4 were inspired by Hiltunen's (2008b) three-dimensional framework of weak signals, this study can give support for their importance for topics, and weak signals. Apart from that, the importance of the last Feature F5 inspired by (Holopainen & Toivonen, 2012) can also support their findings.

Revisiting the study critically, revealed certain bottlenecks. During the data collection and especially the refinement of the dataset by removing unfit samples, a large proportion of the data initially collected was lost. Furthermore taking months as length for the time slices results in a larger number of time slices but poses the possibility of a smaller number of coherent

topics (Greene & Cross, 2017). Also, the final dataset used for binary classification exhibited a skewed class distribution, which was counteracted using performance indicators unaffected by this limitation. Next, effectively the result of an unsupervised machine learning method was used to train a supervised classifier. This approach is in itself novel and not a standard procedure, and therefore maybe error prone.

The findings above, also indicate that the algorithm used for extracting topics tends to find a large number of smaller topics. Whereas optimizing the algorithm by optimizing the coherence measure c_v, the results are very dependent on this measure.

6.3 Impact of the study on weak signal research

In weak signal theory, weak signals are thought of as warnings of possible changes in the future. Their transition to strong signals bears the utmost interest of scholars and practitioners alike. This very transition also poses the central strategic paradox of weak signal theory. Actors can accept incompleteness in the form of weak signals or wait for more complete information to plan upon in the form of strong signals, which is likely too late. Ansoff (1975) proposes detection and monitoring of weak signals as a strategic response to weak signals. This strategic issue management, however, does not solve the paradox, because it is also based on waiting. This study proposes another viable response to weak signals. Instead of monitoring and planning a graduated response, this study introduces forecasting of weak signals as a form of response to weak signals. This constitutes a first attempt of replacing the reliance on weak signals with a forecast of strong signals which could potentially solve this strategic paradox.

This study also has an impact of other areas of weak signal research focusing on the operationalization and application of the concept. Therefore, this study uses web content as new source of data and applies semantic text classification to derive signals. Doing so this study confirms that this data and the method is valid for detecting weak signals. Moreover, originally the detection of weak signals required experts, by applying the three dimensions framework from Hiltunen (2008b) this study shows that this part can be done automatically, which questions the role of experts in weak signal detection in the future.

7 Conclusion

This study proposes a new approach to automatically detect and predict weak signals. The research objective involved an automatic solution working on a continuous basis and independent from manual intervention. This solution is aimed at supporting strategic decision makers about the current environment and give forecasts of weak signal appearance in the future.

This study, by applying this process abstracted information from documents to topic chains, to find underlying structure and insights.

7.1 Theoretical contributions

This paper aimed at understanding and predicting the occurrence of weak signals. Until now multiple scholars have used weak signals in several contexts, but to the knowledge of the author, no one attempted to forecast weak signal evolution and the associated predictors. This study reveals that weak signals indeed can be forecasted with a high precision.

In order to foresee the future with weak signals, the author aimed at predicting the occurrence of weak signals and understand its most important predictors. Previous work in the field of weak signals and especially quantification of such inspired the derivation of features to help predict the occurrence. After the extraction of topics for various time slices from a corpus of date-stamped documents, those topics were compared to find related topics throughout different time slices. Afterward, a binary classifier was trained to learn which topics have or have not consecutive similar topics by using the criteria defined earlier.

This study contributes to theory in multiple ways. First, it confirms that weak signals can be found in web content data. Second, whereas Ansoff (1975) proposed strategic issue management as a response to weak signals, this study shows an alternative response in the form of forecasting weak signals instead of monitoring. Third, the evaluation showed that the concept of weak signals is still valid for practitioners and strategic decision makers alike. Fourth, the results about the most important predictors confirm the two dimensions signal and issue of weak signals proposed by Hiltunen (2008b) are important to forecasting signals.

7.2 Practical and managerial contributions

This work and the derived artifacts help practitioners to scan the environment and to detect and forecast weak signals. Whereas the proposed procedure is completely unsupervised, an automatic tool can be created to continuously mine documents, extract topics, and inform about emerging weak signals and predict their future. Even without predicting the evolution the

proposed artifacts give valuable insights into the topics occurring in the environment of a business. This is a major advancement compared to trend extrapolation or manual expert panels.

Apart from the general contributions, the proposed artifacts also allow the interaction with the data. Unlike static reports, practitioners can set their own set of parameters and investigate the findings. Discussions about the advancements with actual users, revealed the positive cost effects, level of automation, time savings, objectivity and level of insights.

This study also advances technology foresight by proposing a solution of automatic environmental scanning and weak signal detection without the intervention of human experts. This study, therefore, also close a gap and a bottleneck within technology foresight.

7.3 Limitations and future research

The study has multiple limitations. First, the data itself is only gathered from one source, and apart from that web data is unstructured. It contains different sets of information, which can be hard to structure. Second, the data cleaning step is very critical, and automating this step entirely might include unfit samples into the final dataset. Third, changes in the algorithm of Google, the primary data source, might also bring unforeseeable consequences to the whole design. Fourth, the internet itself changed rapidly throughout the last decade. This study hence includes data from arguably another internet era. Fifth, the interpretation of the artifacts is not trivial, and more importantly not automatable. Sixth, the underlying theoretical concept of weak signals is still vague. Moreover, the differentiation towards strong signals is not discussed in details throughout research. The operationalization and measurement of weak and strong signals is weakly discussed in literature.

This initial work on forecasting weak signals focused on the binary prediction of the occurrence of weak signals and weak signal topics. Future work could also aim at predicting the length of weak signal topic chains; therefore, the length in time slices a topic has similar topics. By accomplishing more accurate predictions about weak signals, and estimating their information content, weak signals become strong signals. One of the differences between both phenomena is the probability of realization. Predicting this would enhance strategic planning by identifying long-term business opportunities at an early stage. Furthermore, this design could also be used to incorporate more diverse data sources, like social media data, patent data, or crowdfunding data and combine them. Also, future work could involve a new iteration of the design to involve the feedback of practitioners to sharpen the artifacts. Lastly, this design could also be applied to other areas of interest. Topics and weak signals cannot only be identified from web documents but also from all sorts of documents, future fields of application could, for instance, be incident management in customer service departments.

8 References

Aharonson, B. S., & Schilling, M. A. (2016). Mapping the technological landscape: Measuring technology distance, technological footprints, and technology evolution. *Research Policy*, *45*(1), 81–96. https://doi.org/10.1016/j.respol.2015.08.001

Ansoff, H. I. (1975). Managing Strategic Surprise by Response to Weak Signals. *California Management Review*, *XVIII*(2), 21–33. https://doi.org/citeulike-article-id:1109593

Ansoff, H. I. (1980). Strategic Issue Management. *Strategic Management Journal*, *1*(2), 131–148.

Ansoff, H. I. (1982). *Strategic response in turbulent environments*. European Institute for Advanced Studies in Management Brussels.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77. https://doi.org/10.1145/2133806.2133826

Blei, D. M., & Lafferty, J. D. (2006). Dynamic topic models. *Proceedings of the 23rd International Conference on Machine Learning - ICML '06*, 113–120. https://doi.org/10.1145/1143844.1143859

Blei, D., Ng, A., & Jordan, M. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*(Jan), 993–1022.

Breitzman, A., & Thomas, P. (2015). The Emerging Clusters Model: A tool for identifying emerging technologies across multiple patent systems. *Research Policy*, *44*(1), 195–205. https://doi.org/10.1016/j.respol.2014.06.006

Chang, J., Gerrish, S., Wang, C., & Blei, D. M. (2009). Reading Tea Leaves: How Humans Interpret Topic Models. *Advances in Neural Information Processing Systems 22*, 288--296. https://doi.org/10.1.1.100.1089

Choo, C. W. (2002). *Information management for the intelligent organization: the art of scanning the environment*. Information Today, Inc.

Chuang, J., Manning, C. D., & Heer, J. (2012). Termite : Visualization Techniques for Assessing Textual Topic Models. *Proceedings of the International Working Conference on Advanced Visual Interfaces - AVI '12*, 74. https://doi.org/10.1145/2254556.2254572

Day, G. S., & Schoemaker, P. J. H. (2006). *Peripheral vision: Detecting the weak signals that will make or break your company*. Harvard Business Press.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, *41*(6), 391.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

Greene, D., & Cross, J. P. (2017). Exploring the political agenda of the european parliament using a dynamic topic modeling approach. *Political Analysis*, *25*(1), 77–94. https://doi.org/10.1017/pan.2016.7

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design Science in Information Systems Research. *MIS Quarterly*, *28*(1), 75–105. https://doi.org/10.2307/25148625

Hiltunen, E. (2007). *Where do future oriented people find weak signals?*

Hiltunen, E. (2008a). Good sources of weak signals: A global study of where futurists look for weak signals. *Journal of Futures Studies*, *12*(4), 21–44. https://doi.org/10.1016/j.futures.2011.10.002

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

Hirschberg, J., & Manning, C. D. (2015). Advances in natural language processing. *Science*, *349*(6245), 394–416. https://doi.org/10.1017/CBO9781139051729.024

Holopainen, M., & Toivonen, M. (2012). Weak signals: Ansoff today. *Futures*, *44*(3), 198–205. https://doi.org/10.1016/j.futures.2011.10.002

Horton, A. (1999). A simple guide to successful foresight. *Foresight*, *1*(1), 5–9. https://doi.org/10.1108/14636689910802052

Kayser, V., & Blind, K. (2017). Extending the knowledge base of foresight: The contribution of text mining. *Technological Forecasting and Social Change*, *116*, 208–215. https://doi.org/10.1016/j.techfore.2016.10.017

Keller, J., & von der Gracht, H. A. (2014). The influence of information and communication technology (ICT) on future foresight processes - Results from a Delphi survey. *Technological Forecasting and Social Change*, *85*, 81–92. https://doi.org/10.1016/j.techfore.2013.07.010

Khalili, A., & Auer, S. (2013). WYSIWYM Authoring of Structured Content Based on Schema.org (Vol. 8181, pp. 425–438). https://doi.org/10.1007/978-3-642-41154-0_32

Kim, D., & Oh, A. (2011). Topic chains for understanding a news corpus. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. https://doi.org/10.1007/978-3-642-19437-5_13

Kim, J., Hwang, M., Jeong, D. H., & Jung, H. (2012). Technology trends analysis and forecasting application based on decision tree and statistical feature analysis. *Expert Systems with Applications*, *39*(16), 12618–12625.

https://doi.org/10.1016/j.eswa.2012.05.021

Kolay, S. (2008). A larger scale study of robots.txt. *Proceeding of the 17th International Conference on World Wide Web - WWW '08*, 1171. https://doi.org/10.1145/1367497.1367711

Kotsiantis, S. B. (2007). Supervised Machine Learning: A Review of Classification Techniques. *Informatica*, *31*, 249–268. https://doi.org/10.1115/1.1559160

Loper, E., & Bird, S. (2002). NLTK: The Natural Language Toolkit, 63–70. https://doi.org/10.3115/1118108.1118117

Miles, I., Meissner, D., Vonortas, N. S., & Carayannis, E. (2017). Technology foresight in transition. *Technological Forecasting and Social Change*, *119*(April), 211–218. https://doi.org/10.1016/j.techfore.2017.04.009

Molitor, G. T. T. (2003). Molitor Forecasting Model: Key Dimensions for Plotting the "Patterns of Change." *Journal of Futures Studies*, *8*(1), 61–72.

Mühlroth, C., & Grottke, M. (2018). A systematic literature review of mining weak signals and trends for corporate foresight. *Journal of Business Economics*, *88*(5), 643–687. https://doi.org/10.1007/s11573-018-0898-4

Olson, R. S., Bartley, N., Urbanowicz, R. J., & Moore, J. H. (2016). Evaluation of a Tree-based Pipeline Optimization Tool for Automating Data Science. In *Proceedings of the Genetic and Evolutionary Computation Conference 2016* (pp. 485–492). New York, NY, USA: ACM. https://doi.org/10.1145/2908812.2908918

Olson, R. S., Urbanowicz, R. J., Andrews, P. C., Lavender, N. A., Kidd, L. C., & Moore, J. H. (2016). Applications of Evolutionary Computation: 19th European Conference, EvoApplications 2016, Porto, Portugal, March 30 -- April 1, 2016, Proceedings, Part I. In G. Squillero & P. Burelli (Eds.) (pp. 123–137). Springer International Publishing. https://doi.org/10.1007/978-3-319-31204-0_9

Park, H., Kim, K., Choi, S., & Yoon, J. (2013). A patent intelligence system for strategic technology planning. *Expert Systems with Applications*, *40*(7), 2373–2390. https://doi.org/10.1016/j.eswa.2012.10.073

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A Design Science Research Methodology for Information Systems Research. *Journal of Management Information Systems*, *24*(3), 45–77. https://doi.org/10.2753/MIS0742-1222240302

Pépin, L., Kuntz, P., Blanchard, J., Guillet, F., & Suignard, P. (2017). Computers & Industrial Engineering Visual analytics for exploring topic long-term evolution and detecting weak signals in company targeted tweets, *112*, 450–458.

https://doi.org/10.1016/j.cie.2017.01.025

Pietrobelli, C., & Puppato, F. (2016). Technology foresight and industrial strategy. *Technological Forecasting and Social Change*, *110*, 117–125. https://doi.org/10.1016/j.techfore.2015.10.021

Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the Space of Topic Coherence Measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining - WSDM '15*. https://doi.org/10.1145/2684822.2685324

Sasaki, Y. (2007). The truth of the F-measure. *Teach Tutor Mater*, 1–5. Retrieved from http://www.cs.odu.edu/~mukka/cs795sum09dm/Lecturenotes/Day3/F-measure-YS-26Oct07.pdf

Sievert, C., & Shirley, K. (2014). LDAvis: A method for visualizing and interpreting topics. *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, 63–70. https://doi.org/10.1.1.100.1089

Sulo, R., Berger-Wolf, T., & Grossman, R. (2010). Meaningful selection of temporal resolution for dynamic networks. *Proceedings of the Eighth Workshop on Mining and Learning with Graphs*, 127–136. https://doi.org/10.1145/1830252.1830269

Thorleuchter, D., & Van den Poel, D. (2016). Identification of interdisciplinary ideas. *Information Processing and Management*, *52*(6), 1074–1085. https://doi.org/10.1016/j.ipm.2016.04.010

Thorleuchter, D., & Van Den Poel, D. (2013a). Weak signal identification with semantic web mining. *Expert Systems with Applications*, *40*(12), 4978–4985. https://doi.org/10.1016/j.eswa.2013.03.002

Thorleuchter, D., & Van Den Poel, D. (2013b). Web mining based extraction of problem solution ideas. *Expert Systems with Applications*, *40*(10), 3961–3969. https://doi.org/10.1016/j.eswa.2013.01.013

Thorleuchter, D., Van Den Poel, D., & Prinzie, A. (2010). Mining ideas from textual information. *Expert Systems with Applications*, *37*(10), 7182–7188. https://doi.org/10.1016/j.eswa.2010.04.013

Voros, J. (2003). A generic foresight process framework. *Foresight*, *5*(3), 10–21. https://doi.org/10.1108/14636680310698379

Yoon, J. (2012). Detecting weak signals for long-term business opportunities using text mining of Web news. *Expert Systems with Applications*, *39*(16), 12543–12550. https://doi.org/10.1016/j.eswa.2012.04.059

Appendix

Appendix A: Topic chain emergence map

Topic chain emergence map artifact whereas a selected topic from the map on the left shows the most frequent terms on the right.