Futures 66 (2015) 25–34

Contents lists available at ScienceDirect

Futures

journal homepage: www.elsevier.com/locate/futures

Idea mining for web-based weak signal detection

D. Thorleuchter [a,*], D. Van den Poel [b,1]

[a] Fraunhofer INT, D-53879 Euskirchen, Appelsgarten 2, Germany
[b] Ghent University, B-9000 Gent, Tweekerkenstraat 2, Belgium

ARTICLE INFO

Article history:
Available online 19 December 2014

Keywords:
Web mining
Strategic decision making
Idea mining
Weak signal analysis

ABSTRACT

We investigate the impact of idea mining filtering on web-based weak signal detection to improve strategic decision making. Existing approaches for identifying weak signals in strategic decision making use environmental scanning procedures based on standard filtering algorithms. These algorithms discard patterns with low information content; however, they are not able to discard patterns with low relevance to a given strategic problem. Idea mining is proposed as an algorithm that identifies relevant textual patterns from documents or websites to solve a given (strategic) problem. Thus, it enables to estimate patterns' relevance to the given strategic problem. The provided new methodology that combines weak signal analysis and idea mining is in contrast to existing methodologies. In a case study, a web-based scanning procedure is implemented to identify textual internet data in the field of self-sufficient energy supply. Idea mining is applied for filtering and weak signals are identified based on the proposed approach. The proposed approach is compared to a further – already evaluated – approach processed without using idea mining. The results show that idea mining filtering improves quality of weak signal analysis. This supports decision makers by providing early and suggestive signals of potentially emerging trends, even with only little expressive strength.

1. Introduction

A well-known approach to support strategic planning with forecasting is to identify and to analyze the development of signals (Holopainen & Toivonen, 2012; Rossel, 2009). Ansoff (1975) defines a signal as an existing event. To become relevant, this event should have an impact on at least one target related to the strategic planning decision (Driscoll, 1997; Mizzaro, 1997; Shanteau, 1992). While signals' impact is not static, Ansoff distinguishes between strong and weak signals (Rossel, 2012): Strong signals are 'sufficiently visible and concrete' events while weak signals are 'imprecise early indications about impending impactful events'. Strong signals are normally well-known for decision makers and anyway, they are considered in the decision making process. Weak signals are 'features of incipient changes that can help managers avoid strategic surprises' (Ansoff, 1975), however; they are possibly unknown for decision makers or they are ignored by decision makers and thus, they are not considered in the decision making process (Ilmola & Kuusi, 2006). Evaluating the impact development of these weak signals helps strategic planners to consider future impacts on their strategic decision by time (Mendonça, Cardoso, & Caraça, 2012). This is important because strategic planning is characterized by long-term decision making.

* Corresponding author. Tel.: +49 2251 18305; fax: +49 2251 18 38 305.
E-mail addresses: dirk.thorleuchter@int.fraunhofer.de (D. Thorleuchter), dirk.vandenpoel@ugent.be (D. Van den Poel).
[1] Tel.: +32 9264 8980.

http://dx.doi.org/10.1016/j.futures.2014.12.007
0016-3287/

Related work for evaluating these developments (weak signal analysis approaches) from textual information use standard filtering algorithms from text mining to discard patterns with low information content.

In contrast to related work, a new methodology is provided where idea mining (Thorleuchter, Van den Poel, & Prinzie, 2010) is used for filtering because it enables to select textual patterns based on their relevance to a given organization's target. Thus, it discards both, patterns with low information content and patterns with low relevance to organization's target. The scope of this research is to show an improved performance of the proposed approach compared to existing weak signal analysis approaches.

In a case study, we implement this new methodology to support strategic decision makers in the field of self-sufficient energy supply. Textual information is collected from the internet by an environmental scanning procedure. Ideas are extracted from the collected data using idea mining. The extracted ideas are clustered with latent semantic indexing. Weak signals are identified representing ideas that might be of relevance for strategic decision making. Overall, evaluated results from the case study show that weak signal analysis can be improved by using idea mining.

Weak signal analysis is based on environmental scanning as described in Section 2.1. Existing approaches are introduced in Section 2.2. Some approaches (including our newly introduced approach) use semantic clustering for weak signal analysis as introduced in Section 2.3. A description of idea mining is given in Section 2.4. Different steps of the proposed methodology are depicted in Section 3 and explained in several sub sections. The description of processing a case study can be found in Section 4. It includes a comparison to a different approach for evaluation purpose (see Section 4.4). Examples of case study results are depicted in Section 4.5. Section 5 concludes the paper.

2. Background

2.1. Environmental scanning

For the strategic planning of an organization, relevant weak signals occur in the organization's environment (Ansoff, 1984). To identify these weak signals, a wide scope environmental scanning process is suggested (Kim et al., 2013; Tonn, 2008). After defining organization's environment, it identifies all existing data sources within the environment and it traces their development (Kaivooja, 2012). While today the internet becomes more and more the most important data source for organizations, environmental scanning is often applied on internet information (Decker, Wagner, & Scholz, 2005). Most of this information (websites, internet blogs, news feeds, social media posts, etc.) is accessible in form of textual data (Uskali, 2005) and furthermore, it is named documents.

An internet based environmental scanning collects a large number of documents related to an organization because of the huge number of textual information existing in the internet and the high degree of cross-linkage within the documents (Abebe, Angriawan, & Tran, 2010; Choo & Auster, 1993; Choo, 1999; Liu, Shih, Liau, & Lai, 2009). Each document consists of a specific number of terms (e.g. words) that can be semantically summarized to textual patterns. Thus, analyzing this large number of documents as well as the textual patterns standing behind the documents requires at least the use of a semi-automatic approach (Yang, 2009). This high computational complexity is increased further by the fact that a textual document is often related to several different events (Uskali, 2005).

The identification of events from documents can be done by applying existing text mining approaches. They extract textual patterns from the documents e.g. based on semantic clustering. Each text pattern is investigated in order to identify a relationship to the organization and specifically to the strategic decision problem (e.g. an organization is leading in a specific technology and terms describing this technology can be found in the text pattern). Thus, the corresponding event impacts organization's target and can be seen as a signal (Tabatabei, 2011).

To classify this signal as 'weak signal' or as 'strong signal', it is important to know that Ansoff and McDonnell (1990) define five stages for signal development from weak to strong: '(1) the sense of threat/opportunity, (2) the source of threat/opportunity is known, (3) the shape of threat/opportunity is concrete, (4) the response strategies are understood and (5) the outcome of response is forecastable.'

The semantic clustering shows groups of these selected text patterns (signals) occurring in a large number of different documents. They can be seen as (strong) signals at stage three, four, or five because the signals itself as well as their impact is widely mentioned in the internet. In contrast to this, semantic clustering also shows signals occurring in a small number of internet documents. The documents should be independent of each other to decrease probability that the signal is a hoax. These signals are mentioned by a small number of different document sources and thus, Thorleuchter and Van den Poel (2013a) interpret them as weak signals according to stage two. Based on this interpretation, we are aware that signals widely discussed in the internet but not recognized as a new threat/opportunity are misclassified. Thus, detecting weak signals at stage one is excluded by the proposed methodology.

2.2. Weak signal analysis approaches

In literature, many attempts to realize weak signal analysis according to Ansoff can be seen. Schwarz (2005) proposes a weak signal analysis approach to identify new and arising technologies. These technologies should be relevant for high tech companies in Europe. Unfortunately, applying this theoretical approach fails in practice. This is because of the very high manual effort caused by collecting the required data. While an automated environmental scanning tool was not available for him, human experts have manually scanned the environment of high tech companies by reading newspaper articles, blogs,

internet websites, etc. related to the companies. Despite the very high manual effort, the clustering results were poor (Tabatabei, 2011).

Examples for successfully applied approaches from literature are provided by Decker et al. (2005) and Uskali (2005). Both approaches restrict the number of input data to a very small size to prevent the high manual effort. Thus, only a small number of documents e.g. articles from a specific newspaper are considered. Despite the success of both approaches, they are restricted to small number of input data and they could not be extended to consider a large number of input data as they probably would get by applying a wide scope internet based environmental scanning (Tabatabei, 2011).

Examples for an automated internet based environmental scanning combined with weak signal analysis are provided by Yoon (2012). Yoon identifies trends related to solar cells by restricting the internet based environmental scanning on news articles. Despite interesting results, an evaluation of his proposed approach is missing (Tabatabei, 2011).

Tabatabei (2011) identifies new and arising trends in the field of digital media by applying a wide scope internet based environmental scanning. He uses a knowledge structure based approach for clustering. An extension of his work is proposed by Thorleuchter and Van den Poel (2013a). In contrast to the knowledge structure based approach used by Tabatabei, Yoon, Decker, Uskali, and Schwarz, they provide a semantic clustering approach. Especially by analyzing unstructured texts (as found in internet documents), semantic approaches normally outperform knowledge structure based approaches (Cao, Duan, & Gan, 2011). This is because texts in the internet (websites, blogs, etc.) are written by different persons. Each person has a personal writing style and proposes information in a specific personal context. Semantic approaches focus on the meaning of a text rather than on the used words. This enables to identify similar texts if they share a common meaning regardless whether the texts are written in different writing styles or in different contexts.

Thorleuchter and Van den Poel (2013a) filter results of an internet based environmental scanning with standard text mining filtering methods from raw text cleaning over correcting typographical errors up to stop-word filtering and stemming. The filtered results are semantically clustered. Based on this clustering, weak and strong signals are identified. The relevance of the retrieved results from environmental scanning is considered by selecting and executing several internet search queries related to a given strategic problem. However, an internet search result normally contains irrelevant results, too. Further, each result item possibly consists of several paragraphs and it might be that only one paragraph is relevant for the strategic decision problem while the others are irrelevant. The disadvantage of the approach by Thorleuchter and Van den Poel (2013a) is that all search query results are considered and thus, many irrelevant documents and paragraphs are also used for weak signal analysis.

2.3. Semantic clustering for weak signal analysis

Thorleuchter and Van den Poel (2013a) use semantic clustering for weak signal analysis to consider differences in authors' writing styles and contexts. Semantic approaches calculate dependencies among terms e.g. by using eigenvector techniques from algebra to group semantically related terms (clusters) (Jiang, Berry, Donato, Ostrouchov, & Grady, 1999). Each group consists of terms that occur together in several documents but it also consists of terms that might occur together in these documents. This considers the aspect of meaning from the documents. It is in contrast to knowledge structure based approaches that consider the aspects of words (Kontostathis & Pottenger, 2006).

To identify weak signals from the clustering results, a weak signal maximization (WSM) approach is introduced (Thorleuchter & Van den Poel, 2013a). Semantic clustering is applied several times, each time to create a different number of clusters (k). For each time, a human expert analyzes the k clusters to identify weak signals. If k is too small then several weak signals might be found in one cluster. If k is too large then a weak signal possibly occurs in several clusters. To calculate an optimal value of k, the number of clusters is calculated where one cluster represents one and only one weak signal (one-to-one correspondence). This k is selected where the calculated number is maximal.

Thorleuchter and Van den Poel (2013a) use latent semantic indexing (LSI) for semantic clustering. However, modern approaches are proposed in literature with an improved performance to LSI. Examples are 'Probabilistic Latent Semantic Indexing' (Hofmann, 1999), 'Non-Negative Matrix Factorization' (Lee & Seung, 1999), and 'Latent Dirichlet Allocation' (Blei, Ng, & Jordan, 2003; Ramirez, Brena, Magatti, & Stella, 2012; Xianghua, Guo, Yanyan, & Zhiqiang, 2013).

These modern approaches are not used in the proposed methodology (see Section 3) because of two reasons. First, the proposed approach should be compared to the corresponding approach that uses LSI but that does not use idea mining. Thus, our proposed methodology also uses LSI for semantic clustering. Second, if the use of idea mining for an LSI-based weak signal analysis is successful then it will be successful for a PLSI, NMF, or LDA-based weak signal analysis anyway. This is because PLSI, NMF, or LDA are of better performance than LSI.

2.4. Idea mining

Idea mining (Thorleuchter, Van den Poel, & Prinzie, 2010) aims on the extraction of technological ideas from textual data. The extracted ideas should serve as problem solution ideas for a given task. Idea mining is based on technique philosophy where an idea is defined as a means together with a corresponding end. Means and ends are seen as textual patterns that consist of several technical terms (words) occurring together. Thus, an idea is defined as a textual pattern where terms describing a means and a corresponding end co-occur. To apply idea mining, a user has to provide a text (furthermore named context) where the problem is described. A user also has to provide textual data (furthermore named idea text) where new and useful ideas are described that can be used to solve the given problem.

The context consists of ideas that are known to the users. According to technique philosophy, a known idea consist of a known means and a corresponding known end. Idea mining extracts the terms representing the known means and the known ends from the context.

The idea text consists of unknown ideas from user's point of view that possibly can be used for problem solving. Means and ends from these unknown ideas are identified and the corresponding terms are extracted. Comparing means and ends from idea text and context leads to the identification of new ideas where either a known means occurs together with an unknown end or an unknown means occurs together with a known end. The first case shows that e.g. an existing technology used for a specific application field can also be used in a new application field and the second case shows e.g. a new technology that possibly could replace an existing technology used in a specific application field. Furthermore, terms extracted from known means and ends are named known terms and terms extracted from unknown means and ends are named unknown terms.

Idea mining identifies a textual pattern from the idea text as a new idea (that can be used to solve an existing problem) if 'the number of known terms and the number of unknown terms in this pattern are well balanced, if known terms occur more frequently in context than other terms, if unknown terms occur more frequently in idea text than other terms, and if specific terms occur, which are characteristic for a new idea' (Thorleuchter & Van den Poel, 2013b).

Ideas from different domains have different idea characteristics. This can be considered by selecting parameters: With variable selection (Coussement & Van den Poel, 2008), the parameters are ordered based on their performance impact (as calculated by their χ^2-statistic). With a forward-selection procedure, a specific threshold is determined to select parameters with impact above the threshold (Van den Poel & Buckinx, 2005). The reduced number of parameters enables to calculate an optimized value for each of the selected parameters in order to get the highest cross-validated *F*-measure (Guns, Lioma, & Larsen, 2012). This is realized by applying a grid search (Basrak, 1987; Jiménez, Lázaro, & Dorronsoro, 2009) where discrete sequences are used to set the parameter values.

3. Methodology

3.1. Overview

The proposed approach is depicted in Fig. 1. It has the aim to identify new, arising signals from organization's environment with relevance to a given strategic decision problem of the organization. Thus, a description of a strategic decision problem should be provided in plain text form. Key words should be comprised in the text and the problem should be addressed clearly and comprehensible. This input data is processed in several steps. A preprocessing and term vector creating step (see Section 3.2) is applied twice in the approach. It transforms the data received from the previous step to a specific format as required by the subsequent step. Standard methods from text mining are used in this step. A web mining step (see Section 3.3) is applied to identify textual information from the internet that is related to the given strategic decision problem. It creates and executes search queries based on the given key words and it crawls the full text of the retrieved results. As a result, a document collection of the internet representing organization's environment is given. The idea mining step (see Section 3.4) compares the problem description to the document collection resulted by the web mining step. It identifies new ideas within the collection that might be able to solve the given strategic decision problem. Based on the identified ideas, a weak signal analysis step (see Section 3.5) is processed. It clusters the new ideas, it identifies weak signals, and it provides weak signals to the decision maker for an improved strategic decision making. The overall performance of the proposed methodology is calculated in an evaluation step. Standard performance measures are used to show the precision and the recall of the results based on a manual evaluation by human experts. Further, a comparison to the results of a related approach processed without idea mining is done. This shows the impact of idea mining on the proposed methodology.

3.2. Preprocessing and term vector creating

Preprocessing is an important factor in processing unstructured texts. Standard methods from text mining are used to identify key words from a given text: Raw text cleaning deletes images, scripting code, specific characters, and punctuation from the text. Tokenization splits unstructured texts in terms whereby the term unit normally is defined as words. Typo correction compares terms to a dictionary to identify typographical errors and to correct them. Case conversion ensures that all terms are written in the same way e.g. with a capitalized first sign and with all further signs in lower case.

Further methods are used to filter information by grouping related terms or by discarding low-informative terms: Stop word filtering identifies terms that frequently occur in all documents of a collection because they only bear little content information. Part-of-speech tagging enables to identify the grammatical role of a term in a group of words. Based on the specific role, its informative value can be estimated. Zipf's law (Zeng, Duan, Cao, & Wu, 2012; Zipf, 1949) discards the large number of terms that appear only once or twice because it is estimated that these terms also are low-informative. Stemming uses production rules to group related terms based on their stems.

Term vector creating in vector space model is a standard method for a formal representation of unstructured texts. This method is based on the preprocessing methods as describe above. Usually, a term vector is created for a document.

Term vectors (Nasir, Varlamis, Karim, & Tsatsaronis, 2013) are created on one hand from the description of the strategic decision problem and on the other hand from the results of web mining. The description of the strategic decision problem

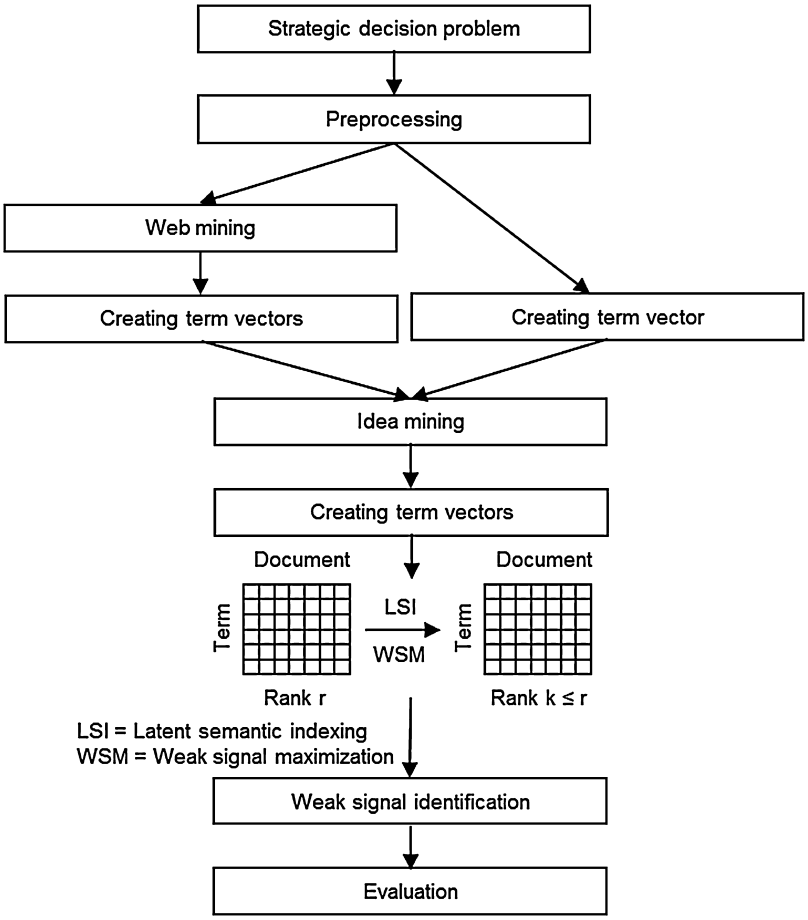

Fig. 1. Based on an existing strategic decision problem, key words describing the problem are identified, composed to several internet search queries, and executed to obtain internet documents. Idea mining identifies textual patterns occurring within the documents that might be a new idea with relevance to the strategic decision problem. A collection of these patterns is used to build a term-document matrix. LSI with WSM is used to reduce the matrix rank from r to k. Human experts identify the weak signals based on the results. An evaluation is done to show the overall success of this approach and to compare the approach to a further weak signal approach processed without idea mining.

contains existing ideas already applied within the organization. They can be found in the description because strategic decision makers assume that new ideas will be more successful in future than the existing ideas. Thus, the low-performance of existing ideas causes the strategic decision problem and the term vector created from this description represents existing ideas. Term vectors created from the resulting documents of web mining also represent existing ideas described in the internet if the term vector is similar to its corresponding term vector from the strategic problem description. However, they represent new ideas, too. This is because many new solution ideas can be found in the internet. Both kinds of term vectors are used for idea mining to identify new ideas related to the given strategic decision problem and to distinguish these ideas to existing ideas.

3.3. Web mining

Web content mining is used to identify and to crawl relevant documents (e.g. websites, blogs, etc.) from the internet. Relevant documents are documents that are related to the strategic decision problem e.g. where new ideas are described that might be useful for consideration in strategic decision making. The results are transformed to term vectors (see Section 3.2).

Documents are identified by use of an internet search engine. This requires the creation of search queries that represent (parts of) the strategic decision problem. While a strategic decision problem is often complex, it consists of different aspects. It is not possible to address all aspects with only one search query. Thus, several search queries are created that address all aspects in total. The content of the search queries is based on the key words extracted from the given strategic decision problem (see Section 3.2). These terms are combined – by considering their co-occurrences – to sets of terms each consists of three to five terms (Thorleuchter & Van den Poel, 2013b).

To execute the created search queries, internet search engines are used. They provide web services where an automated query execution can be realized using the advanced programming interfaces (Thorleuchter & Van den Poel, 2013b). Besides title and short description, the query results consist of a hyperlink for each retrieved document. The hyperlinks are collected in a list by discarding double entries, a self-developed web crawled is used to extract the full text from the corresponding internet documents, and the pre-processing step is applied to create a term vector for each document.

3.4. Idea mining

Idea mining (see Section 2.2) is applied for each internet document separately. Each document might contain several ideas. Based on the parameters, text patterns are identified that represent the relevant ideas for the strategic decision problem. Terms in the document that do not belong to an identified text pattern are discarded. This reduces the size of the documents. Thus, each document only consists of ideas stemmed from a specific website. A term vector is created for each of these documents.

Idea mining depends on the used domain because the characteristics of ideas in different domains are different (Thorleuchter & Van den Poel, 2013b). Domain dependencies are considered in idea mining by using different parameter values. Thorleuchter & Van den Poel, 2013b have provided different sets of parameter values for several domains. These domains are selected that are related to the strategic decision problem. The corresponding sets of parameter values can be used to build new parameter values specifically for the strategic decision problem. This enables to get parameter values of good quality with low manual effort.

3.5. Latent semantic indexing

The created term vectors from idea mining are used to build a term-by-document matrix A with rank r ($r \leq \min(m,n)$) where m is the number of distinct terms and n the number of documents. Despite the reduced size of the documents, the matrix consist of a large number of components and thus, of a large dimensionality. Further, many values of matrix's components are zero because terms are not equally distributed over the documents. Thus, the rank of the matrix might be lower than its dimensionality. Singular value decomposition (SVD) uses these zero values to reduce matrix's dimensionality (Chen, Chu, & Chen, 2010). Deerwester, Dumais, Furnas, Landauer, and Harshman (1990) show that a reduced matrix dimensionality summarizes terms semantically based on the term co-occurrences in the document collection. This enables a semantic clustering because groups of semantically related terms can be identified and delimited to other groups. The number of clusters is defined as k. It is calculated by the weak signal maximization approach as introduced in Section 2.3.

A well-known technique is this field is LSI that uses SVD to cluster the groups based on their discriminatory power. The term-by-document matrix A is split in tree matrices. Formula (1) shows the calculation

$$A = U \sum V_t \tag{1}$$

Σ is a diagonal $(r \times r)$ matrix that consists of singular values in descending order. The rank of matrix Σ is reduced from r to k by discarding the singular values of Σ from $k+1$ on. Further, matrix U and V are reduced to matrix U_k and $V_k{}^t$ of rank k. This is also done by discarding all columns of U and V^t from $k+1$ on. As a result, a new matrix A_k is built based on the three reduced matrices. It can be seen as an approximation of A with lower rank k (see Formula (2)).

$$A \approx A_k = U_k \, \Sigma_k \, V_k^t \tag{2}$$

The columns of $(m \times k)$ matrix U_k indicate the impact of each term from the documents on each of the calculated clusters. The impact of the documents on the clusters is depicted by the columns of the $(n \times k)$ matrix V_k.

3.6. Weak signal analysis

LSI leads to the creation of k clusters each consists of semantic textual patterns occurring within the document collection. To identify relevant weak signals from the clusters, a definition from Thorleuchter and Van den Poel (2013a) is used. Relevant weak signals are defined as low frequently occurred semantic textual patterns within the document collection where the impact of the patterns on a given hypothesis (here: a strategic decision problem) is above a specific threshold.

For each of the k clusters, matrix V_k is used to identify the number of documents with impact on the corresponding cluster above a threshold. This enables the identification of low frequently occurred semantic textual patterns within the document collection. These clusters are selected while the others are discarded. For each of the selected cluster, matrix U_k is used to identify the terms that represent the corresponding cluster. As a result, terms with an impact value on the corresponding cluster above a threshold are selected. A term vector is built based on the selected terms and compared to the term vector from the strategic decision problem by a standard similarity measure e.g. Jaccard's coefficient. A result value above a specific threshold shows that the corresponding cluster is related to the strategic decision problem and thus, that the cluster can be seen as relevant weak signal based on the definition as mentioned above.

4. Case study

4.1. Overview

In a case study, we support a German governmental organization by identifying weak signals in the field of self-sufficient energy supply from renewable sources. Self-sufficient energy supply becomes more and more interesting for governments. An example is the conflict between Russia and Ukraine in 2014 where many European governments are afraid of potential unreliability of Russian gas delivery caused by this conflict. In recent times, national governments especially those with low energy resources (natural gas, fossil oil, etc.) often try to get independent from energy import for a specific span of time by provisioning a large amount of natural/fossil energy resources. This cost-expensive solution should be reconsidered because of the increased usage of renewable energy sources. Thus, a self-sufficient energy supply based on renewable energy carrier for a country and its inhabitants become more and more a strategic aim for many governments.

To support a German governmental organization by its strategic decision making (weak) signals are identified from the internet. They describe currently used and upcoming techniques in this field that reach e.g. from power-heat coupling and heat pumps over photovoltaic, water, and wind power plant up to biomass combustion. While most of these (renewable energy) techniques lead to a fluctuating generation of energy, short-term and long-term storage techniques are also important signals for self-sufficiency.

This case study identifies signals and specifically weak signals in the field of self-sufficient energy supply to support the strategic decision in which techniques German government should invest. Based on these results, decision makers are aware of existing techniques as well as of new and arising techniques with future impact on self-sufficient energy supply.

4.2. Web mining

A description of the strategic problem is used as starting point for web mining. Key words are extracted using the methods from pre-processing (see Section 3.2). The key words are composed to eight search queries according to Section 3.3: 'Self-sufficient energy supply', 'Self-sufficiency renewable energy', 'Renewable energy system', 'Self-sufficiency home', 'Independent energy supply', 'Off grid system', and 'Self-sufficient components'. They describe self-sufficiency in the energy field and they enable to identify documents in the internet that are related to this topic. The search queries are executed both in English and in German language. This is because Germany is worldwide leading nation in the renewable energy field and English is worldwide most used language. This selection covers current trends and developments as well as different regions worldwide.

A self-developed program is used to send the queries of English and German language via Google search advanced programming interface (API) to the search engine. Each result item consists of a hyperlink. The hyperlinks are collected while title, abstract, and further information is discarded. The list of hyperlinks consists of many double entries. This is because different search queries may lead to the same document. The double entries are deleted. As a result, 5678 hyperlinks are obtained from the search queries.

While each hyperlink represents a corresponding document, a self-developed internet crawler opens each hyperlink, crawls the corresponding full text, and stores this as a document. The crawler also identifies the language of the document. This is done by extracting stop words (see Section 3.2) from each document and by comparing it to a list of English and German stop words. Documents that only contain English stop words are assigned to English while documents that only consist of German stop words (German text) or that consist of English and German stop words (multi-lingual text) are assigned to German language. As a result, a document collection for both, English and German documents is created. It consists of 2629 German documents with size of 18.7 MB and of 3049 English documents with size of 24.9 MB.

To enable a cross-linguistic LSI analysis for documents in English and in German, German documents are translated to English. A self-developed program provides the German documents to Google translate API. The quality of an automated translation is generally poor however, it is sufficient for this task because for further LSI processing, it is rather more important to translate the single words one-to-one than to provide a grammatically right translation. A further aspect is that translating a German document to English reduces document's size. Thus, the overall size of 18.7 MB is reduced to 16.8 MB. The characteristics of the data concerning web mining are depicted in Table 1.

4.3. Idea mining

For applying idea mining, a term vector has to be provided that represents the strategic decision problem. Among others, the term vector contains words used for web mining. Thus, the term vector normally is similar to some term vectors created

Table 1
Characteristics of the data concerning web mining.

Number of search queries	8
Number/size of English documents	3049/24.9 MB
Number/size of German documents	2629/18.7 MB
Number/size of documents after translation	5678/41.7 MB

from parts of the retrieved documents. This enables to identify new ideas within the documents where the corresponding term vector also contains terms that are not in the description of the strategic decision problem (see Section 3.4).

The parameters of idea mining are determined by the case study results of Thorleuchter and Van den Poel (2013b) where an optimized parameter set for the domain 'Propulsion and Power plants' and for the domain 'Electronic Warfare and Directed Energy Technologies' is calculated. Both domains use energy for a specific aim and thus, they are related to the domain standing behind the self-sufficient energy supply. The mean values of the parameters from both domains are used as parameters for the case study.

Processing of idea mining is done by using the algorithm of the internet tool 'Technological Idea Miner', version 1.5 and by applying it on the retrieved documents from web mining. Each document is split in several text patterns. For each text pattern, a probability is given by the tool that the corresponding text pattern might represent a new and useful idea. Text patterns with a probability above a specific threshold are used for further processing. This reduces the size of each document. As a result, the overall size of the 5678 retrieved documents is strongly reduced from 41.7 MB to 4.8 MB.

4.4. Weak signal analysis and evaluation

SAS 9.3 Textminer is used to apply LSI on the collected data. While the selection of k is critical for applying LSI, a rank-k-model is built for each k according to the weak signal maximization approach. The number of one-to-one correspondences is calculated for each k that represents the number of weak signals. This k is selected where the corresponding number is the overall largest number of weak signals. Fig. 2 shows the relationship between these numbers on the y-axis and the value of k on the x-axis. This weak signal maximization approach is applied twice: for the processed data including the idea mining step (proposed approach) and for the data processed without idea mining (compared approach).

As a result, 10 weak signals can be identified by both, the proposed approach and the compared approach. This shows that idea mining filtering does not discard relevant data despite the strongly reduced size of data in the proposed approach. This also shows that recall measure is equal in both approaches. However, differences can be seen concerning the precision measure. While the size of data is much smaller, the proposed approach gets the overall largest number of weak signals at $k = 32$. This is in contrast to the $k = 48$ clusters that LSI has to build to get 10 clusters representing weak signals. A human expert has to check 32 (48) cluster results (each cluster consists of a list of words) to identify the 10 weak signals. This leads to a precision value of 31% for the proposed approach and to 21% for the compared approach at the same recall value. Thus, the proposed approach outperforms the compared approach.

The proposed approach can be seen as successfully evaluated because the compared approach is already successfully evaluated (precision and recall outperform the baseline, see Thorleuchter and Van den Poel, 2014) and it is shown that the proposed approach outperforms the compared approach.

The slope difference in both approaches can be explained by the fact that the compared approach has a larger size of data than the proposed approach and thus, it is less sensitive concerning varying the number of k. Thus, the slope of the compared approach is normally smaller than the slope of the proposed approach. Examples of results especially those that are useful for strategic decision making are depicted in Section 4.5.

4.5. Results

Weak and strong signals identified by the proposed approach are listened below in brief. It is important to know that web is dynamic and processing the approach several months later will lead to different web crawling data and after analyzing the data and extracting conclusions, it will lead to the identification of different weak and strong signals.

Some signals deals with generation of electrical energy from hydro power where run-of-river and marine current power stations are indicated as strong signals and wave and osmotic power stations are indicated as weak signals. Further signals are based on electrical energy generation from wind energy where a strong signal describes horizontal wind turbines and a weak signal describes vertical wind turbines. Using solar energy is a further signal cluster that consists of solar thermal and photovoltaic systems. These signals are indicated as strong signal. A further cluster of signals is the use of biomass for generation of energy. In this context, wood pellets represent a strong signal while methane and hydrogen represent weak signals.

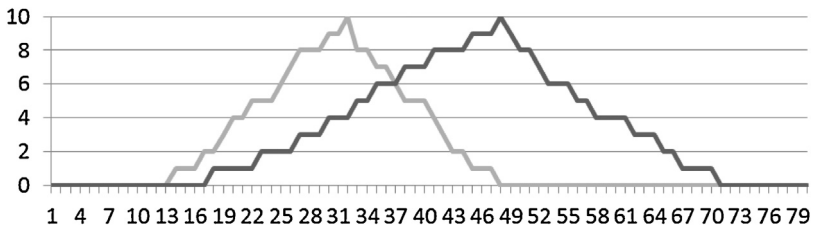

Fig. 2. Number of one-to-one correspondences (y-axis) based on the value of k (x-axis) for the proposed approach (light-gray color) and for the compared approach (dark-gray color).

Several signals also occur in the field of energy storage. In case of electrochemical storage, redox-flow battery is described by a weak signal and lithium batteries as well as sodium-sulphur batteries are described by strong signals. In case of chemical energy storage, hydrogen technology is a weak signal. For the storage of thermal energy, sensible and latent heat storages are strong signals while thermochemical heat storage is a weak signal.

Several signals refer to system-based self-sufficiency. An example for a weak signal is the concept of permaculture. A further weak signal is based on energy-self-sufficient home e.g. passive, low energy, or plus energy house. Strong signals in this area are decentralized energy supply, back-up systems (e.g. emergency generator), and cogeneration of heat and power.

The identified weak signals (wave and osmotic power stations, vertical wind turbine, use of methane and hydrogen in biomass context, redox-flow battery, hydrogen based chemical energy storage, thermochemical heat storage, permacultures, energy-self-sufficient home, etc.) are compared to results of current literature concerning trends in self-sufficient energy supply (Brinkhaus, Jarosch, & Kapischke, 2011; Feroldi, Degliuomini, & Basualdo, 2013; Müller, Stämpfli, Dold, & Hammer, 2011). As a result, literature confirms that these signals represent new and upcoming technological areas where much research is done today to get advances in these areas in future. Thus, a strategic decision maker has to consider these developments today to advance long-term strategic planning.

5. Conclusion

This paper provides a new methodology that combines weak signal analysis with idea mining. The use of idea mining for filtering is in contrast to related approaches. The evaluation considers this fact by comparing the proposed approach to the same approach applied without idea mining as already evaluated. Results show that the provided methodology allows strategic decision makers to consider weak signals for an improved decision making.

In detail, an environmental scanning procedure is applied by web mining to select relevant information from the internet for a strategic problem. Idea mining is used to filter the results and weak signal analysis (LSI and WSM) is used to identify relevant weak signals. The methodology is applied in a case study. Case study results show weak and strong signals in self-sufficient energy supply obtained with better performance than obtained by the approach without idea mining.

While weak signals are context dependent, the results of the case study are often interpretable and disputable. As an example, an energy-self-sufficient home is identified as a weak signal in the context of 'self-sufficient energy supply from renewable sources'. This is because it is seldom mentioned in corresponding internet documents. However, many passive or low energy houses exist today. Thus, it might be a strong signal. A further example is that photovoltaic systems are identified as strong signal in the same context. Considering as background knowledge that many photovoltaic systems are manufactured with very low environmental standards e.g. in China, it should not be a signal in this self-sufficiency context. This example also shows a further limit of the proposed methodology. Weak or strong signals that are not mentioned in the text corpus explicitly (e.g. low environmental standards of solar energy panel production) are not recognized although they directly influence the topic (self-sufficiency). We are aware of the limits of the proposed methodology. Thus for future work, we distinguish between efforts for increasing efficiency of the proposed methodology and efforts for increasing effectivity by extending the methodology with further cross-cutting relations.

To increase efficiency, future work should focus on performance improvement of the proposed methodology by using further well-known methods from semantic clustering e.g. PLSI, NMF, or LDA instead of using LSI. The aim is to show that idea mining filtering also improves quality of PLSI, NMF, or LDA results. A further avenue of future research is to apply the proposed methodology on time series data. This probably enables to improve the tracing of weak signals over time.

In weak signal analysis, the concept of filters especially their impact on weak signal scanning is widely investigated. These filtering impacts on the proposed methodology are not studied in this paper. Thus, they can be studied in future work to extend the methodology. Further, the proposed methodology does not detect weak signals at stage one or wild cards. Future work should extend the methodology to bridge this gap.

References

Abebe, M., Angriawan, A., & Tran, H. (2010). Chief executive external network ties and environmental scanning activities: An empirical examination. *Strategic Management Review, 4*(1), 30–43.

Ansoff, I. H. (1975). Managing strategic surprise by response to weak signals. *California Management Review, 18*(2), 21–33.

Ansoff, I. H. (1984). *Implanting strategic management*. New Jersey: Prentice Hall.

Ansoff, I. H., & McDonnell, E. (1990). *Implanting strategic management* (2nd ed.). New Jersey: Prentice Hall.

Basrak, Z. (1987). A routine for parameter optimization using an accelerated grid-search method. *Computer Physics Communications, 46*(1), 149–154.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research, 3*(4–5), 993–1022.

Brinkhaus, M., Jarosch, D., & Kapischke, J. (2011). All year power supply with of-grid photovoltaic system and clean seasonal power storage. *Solar Energy, 85*(10), 2488–2496.

Cao, Q., Duan, W., & Gan, Q. (2011). Exploring determinants of voting for the helpfulness of online user reviews: A text mining approach. *Decision Support Systems, 50*(2), 511–521.

Chen, M.-Y., Chu, H.-C., & Chen, Y.-M. (2010). Developing a semantic-enable information retrieval mechanism. *Expert Systems with Applications, 37*(1), 322–340.

Choo, C. W. (1999). The art of scanning the environment. *Bulletin of the American Society for Information Science and Technology, 25*(3), 21–24.

Choo, C. W., & Auster, E. (1993). Environmental scanning: Acquisition and use of information by managers. *Annual Review of Information Science and Technology, 28*, 279–314.

Coussement, K., & Van den Poel, D. (2008). Integrating the voice of customers through call center emails into a decision support system for churn prediction. *Information & Management, 45*(3), 164–174.

Decker, R., Wagner, R., & Scholz, S. W. (2005). An internet-based approach to environmental scanning in marketing planning. *Marketing Intelligence & Planning, 23*(2), 189–200.
Deerwester, S., Dumais, S., Furnas, G., Landauer, T., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science, 41*(6), 391–407.
Driscoll, J. R. (1997). Method and system for searching for relevant documents from a text database collection, using statistical ranking, relevancy feedback and small pieces of text. *Laboratory Automation & Information Management, 33*(2), 150.
Feroldi, D., Degliuomini, L. M., & Basualdo, M. (2013). Energy management of a hybrid system on wind-solar power sources and bioethanol. *Chemical Engineering Research and Design, 91*(8), 1440–1455.
Guns, R., Lioma, C., & Larsen, B. (2012). The tipping point: F-score as a function of the number of retrieved items. *Information Processing & Management, 48*(6), 1171–1180.
Hofmann, T. (1999). Probabilistic latent semantic indexing. *Proceedings of the twenty-second annual international SIGIR conference on research and development in information retrieval (SIGIR-99)*.
Holopainen, M., & Toivonen, M. (2012). Weak signals: Ansoff today. *Futures, 44*(3), 198–205.
Ilmola, L., & Kuusi, O. (2006). Filters of weak signals hinder foresight: Monitoring weak signals efficiently in corporate decision-making. *Futures, 38*(8), 908–924.
Jiang, J., Berry, M. W., Donato, J. M., Ostrouchov, G., & Grady, N. W. (1999). Mining consumer product data via latent semantic indexing. *Intelligent Data Analysis, 3*(5), 377–398.
Jiménez, A. B., Lázaro, J. L., & Dorronsoro, J. R. (2009). Finding optimal model parameters by deterministic and annealed focused grid search. *Neurocomputing, 72*(13–15), 2824–2832.
Kaivooja, J. (2012). Weak signals analysis, knowledge management theory and systemic socio-cultural transitions. *Futures, 44*(3), 206–217.
Kim, S., Kim, Y. E., Bae, K. J., Choi, S. B., Park, J. K., Koo, Y. D., Park, Y. W., Choi, H. K., Kang, H. M., & Hong, S. W. (2013). NEST: A quantitative model for detecting emerging trends using a global monitoring expert network and Bayesian network. *Futures, 52*, 59–73.
Kontostathis, A., & Pottenger, W. M. (2006). A framework for understanding latent semantic indexing (LSI) performance. *Information Processing & Management, 42*(1), 56–73.
Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature, 401*(6755), 788–791.
Liu, D. R., Shih, M. J., Liau, C. J., & Lai, C. H. (2009). Mining the change of event trends for decision support in environmental scanning. *Expert Systems with Applications, 36*(2 (Part 1)), 972–984.
Mendonça, S., Cardoso, G., & Caraça, J. (2012). The strategic strength of weak signal analysis. *Futures, 44*(3), 218–228.
Mizzaro, S. (1997). Relevance: The whole history. *Journal of the American Society for Information Science, 48*(9), 810–832.
Müller, M. O., Stämpfli, A., Dold, U., & Hammer, T. (2011). Energy autarky: A conceptual framework for sustainable regional development. *Energy Policy, 39*(10), 5800–5810.
Nasir, J. A., Varlamis, I., Karim, A., & Tsatsaronis, G. (2013). Semantic smoothing for text clustering. *Knowledge-Based Systems, 54*, 216–229.
Ramirez, E. H., Brena, R. F., Magatti, D., & Stella, F. (2012). Topic model validation. *Neurocomputing, 76*(1), 125–133.
Rossel, P. (2009). Weak signals as a flexible framing space for enhanced management and decision-making. *Technology Analysis & Strategic Management, 21*(3), 291–305.
Rossel, P. (2012). Early detection, warnings, weak signals and seeds of change: A turbulent domain of futures studies. *Futures, 44*(3), 229–239.
Schwarz, J. O. (2005). Pitfalls in implementing a strategic early warning system. *Future Studies, 7*(4), 22–31.
Shanteau, J. (1992). How much information does an expert use? Is it relevant?. *Acta Psychologica, 81*(1), 75–86.
Tabatabei, N. (2011). *Detecting weak signals by internet-based environmental scanning* (Master thesis) Waterloo: Waterloo University.
Thorleuchter, D., Van den Poel, D., & Prinzie, A. (2010). Mining Ideas from Textual Information. *Expert Systems with Applications, 37*(10), 7182–7188.
Thorleuchter, D., & Van den Poel, D. (2013a). Weak signal identification with semantic web mining. *Expert Systems with Applications, 40*(12), 4978–4985.
Thorleuchter, D., & Van den Poel, D. (2013b). Web Mining based Extraction of Problem Solution Ideas. *Expert Systems with Applications, 40*(10), 3961–3969.
Thorleuchter, D., & Van den Poel, D. (2014). Semantic Weak Signal Tracing. *Expert Systems with Applications, 41*(11), 5009–5016.
Tonn, B. E. (2008). A methodology for organizing and quantifying the results of environmental scanning exercises. *Technological Forecasting and Social Change, 75*(5), 595–609.
Uskali, T. (2005). Paying attention to weak signals: The key concept for innovation journalism. *Innovation Journalism, 2*(11), 19.
Van den Poel, D., & Buckinx, W. (2005). Predicting Online-Purchasing Behavior. *European Journal of Operational Research, 166*(2), 557–575.
Xianghua, F., Guo, L., Yanyan, G., & Zhiqiang, W. (2013). Multi-aspect sentiment analysis for Chinese online social reviews based on topic modeling and HowNet lexicon. *Knowledge-Based Systems, 37*, 186–195.
Yang, H. C. (2009). Automatic generation of semantically enriched web pages by a text mining approach. *Expert Systems with Applications, 36*(6), 9709–9718.
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.
Zeng, J., Duan, J., Cao, W., & Wu, C. (2012). Topics modeling based on selective Zipf distribution. *Expert Systems with Applications, 39*(7), 6541–6546.
Zipf, G. K. (1949). *Human behaviour and the principle of least effort*. Cambridge: Addison-Wesley.