Expert Systems with Applications 41 (2014) 5009–5016

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Semantic weak signal tracing

Dirk Thorleuchter [a,*], Tobias Scheja [a], Dirk Van den Poel [b]

[a] Fraunhofer INT, Appelsgarten 2, D-53879 Euskirchen, Germany
[b] Ghent University, Faculty of Economics and Business Administration, Tweekerkenstraat 2, B-9000 Gent, Belgium

ARTICLE INFO

Keywords:
Time series
Trend identification
Latent semantic indexing
Web mining

ABSTRACT

The weak signal concept according to Ansoff has the aim to advance strategic early warning. It enables to predict the appearance of events in advance that are relevant for an organization. An example is to predict the appearance of a new and relevant technology for a research organization. Existing approaches detect weak signals based on an environmental scanning procedure that considers textual information from the internet. This is because about 80% of all data in the internet are textual information. The texts are processed by a specific clustering approach where clusters that represent weak signals are identified. In contrast to these related approaches, we propose a new methodology that investigates a sequence of clusters measured at successive points in time. This enables to trace the development of weak signals over time and thus, it enables to identify relevant weak signal developments for organization's decision making in strategic early warning environment.

1. Introduction

Strategic planning for an organization can be improved by the identification and use of signals (Ansoff, 1975). A signal is defined as an event with impact on a specific target or direction. While strategic planning aims on defining directions, signals have to be considered during strategic decision making process. Weak signals are signals where the impact cannot be estimated accurately (Ansoff, 1984). It might be that a new event will possibly have an impact on a target in future. It also might be that an existing event – that does not have an impact on the target up to now – will possibly have an impact in future. For strategic planning, it is hard to identify weak signals from the large number of existing signals. Literature proposes methodologies for weak signal identification. They can be used to identify the future impact of weak signals on own strategic directions.

Weak signals cannot be found in the core area of an organization. This is because all internal events of an organization and their impacts normally are already known by strategic decision makers. Thus, Ansoff shows that weak signals can be found in organization's environment. This requires the use of an environmental scanning procedure to identify signals in a first step. This also requires the use of a clustering approach to group the large number of identified signals and to identify clusters of weak signals in a second step.

The concept of environmental scanning (Tonn, 2008) aims at extracting and analyzing information from various data sources existing in the organization's environment. After analyzing, events can be identified as well as their relationships. Today, the internet is a large and valuable source of information (Decker, Wagner, & Scholz, 2005) where many signals occur. Further, the internet can be used to represent organization's environment. Additionally, most of the data available in the internet are textual data, e.g. websites or blogs. As a result, existing weak signal identification approaches use an environmental scanning that considers textual information from the internet (Decker et al., 2005; Uskali, 2005).

With an internet based environmental scanning, documents e.g. webpages can be identified. This scanning normally has a wide scope and thus, it leads to a large number of extracted internet documents. This makes the use of a (semi-) automatic approach more appropriate than the use of a manual approach. The documents possibly contain texts related to several different topics. Thus, a document as a whole normally does not represent a signal but specific textual patterns that occur within the document probably do (Uskali, 2005). Text mining can be used to extract textual patterns from the full text of the documents and a specific clustering approach can be applied to identify groups of textual patterns that represent weak signals (Tabatabei, 2011; Thorleuchter & Van den Poel, 2013a).

Literature shows some approaches that use internet based environmental scanning for weak signal identification. The approach of Schwarz (2005) aims at the identification of new arising technologies with relevance for the high tech companies in Europe. Unfortunately, the approach could not be applied in practice. It has caused a very high manual effort because an automated environmental

* 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).
URL: http://www.crm.UGent.be (D. Van den Poel).

http://dx.doi.org/10.1016/j.eswa.2014.02.046
0957-4174/

scanning tool was not available and thus, the scanning was processed by human experts. Further, the results of the clustering approach are of low quality. In contrast to this, the approaches of Decker et al. (2005) and Uskali (2005) have been applied successfully. However, they prevent the high manual effort by restricting the number of retrieved documents to a small value. Thus, they could not be seen as wide scope internet based environmental scanning approaches. Tabatabei (2011) provides an automated approach for internet based environmental scanning and clustering. A further knowledge structure based approach is provided by Yoon (2012) that detects weak signal from internet news related to solar cells. Both approaches are applied in a case study however, they are not evaluated. Thorleuchter and Van den Poel (2013a) proposes a semantic clustering approach that can be used together with internet based environmental scanning to identify weak signals. The strength of this semantic approach is that it considers weak signals where the corresponding text patterns are written by different persons, in different writing styles, and in different contexts. Text patterns are recognized as similar if they share a common meaning even if they do not share common words. Based on a clustering of these text patterns, signals can be identified and distinguished in weak and strong signals. The authors prove the feasibility of the proposed approach by evaluating results of a case study.

In contrast to related work, we provide a methodology that investigates a sequence of clusters that represent weak signals. This enables to trace the development of weak signals over time. In a first step, an internet based environmental scanning is processed and semantic clustering is applied to identify signals. This first step is processed in accordance to the methodology proposed by Thorleuchter and Van den Poel (2013a). In a second step, the internet based environmental scanning is repeated at successive points in time. For each point in time, the scanning results are projected into the semantic space created by the clustering approach in the first step. This allows tracing the identified signals. Examples are the identification of weak signals that lose their impact on a target, weak signals that become strong signals with large impact on a target, or weak signals that do not change its impact over time. For clustering and classification, latent semantic indexing (LSI) is used. It enables the identification of semantic textual patterns from large document collection and it also enables clustering and the assignment of new documents to an existing semantic space.

In a case study, the proposed methodology is applied in the field of storages technologies for intermittent energy sources where the development of weak signals is traced. The aim of the case study is to show the general feasibility of the approach. For evaluation, hypotheses about future development of storages technologies were provided by a literature review of future studies. They are compared to the traced weak signals. As a result, the proposed methodology enables to identify future technological developments based on current internet information about energy storage technologies. This supports decision makers by their strategic decision making.

Overall, we propose a methodology that enables to trace the development of weak signals over time. It also considers aspects of meaning from the documents identified by an internet based environmental scanning. Tracing weak signals can support strategic decision making because events with impact on strategic aims or directions can be identified in advance. This allows decision makers to react ahead of time.

2. Background

2.1. Internet based environmental tracing

A huge amount of information can be found in the internet dealing with different topics. Using this information together with traditional information sources (e.g. organization internal databases) provides an added value for decision making (D'Haen, Van den Poel, & Thorleuchter, 2013). An example for using this information for organization's strategic planning is to collect and analyze information from the internet about organization's customers and competitive organizations (Teo & Choo, 2001). The huge amount of information available in the internet enforces the use of web mining approaches. This enables to collect information based on an automated process for scanning all relevant internet websites (Kosala & Blockeel, 2000). Web mining includes the identification of website's relevance for a specific topic and it also includes the process of reducing information from relevant websites (Velásquez, Dujovne, & L'Huillier, 2011). The identification of website's relevance is normally realized by using advanced programming interfaces (APIs) of internet search engines (Thorleuchter & Van den Poel, 2013b). To reduce information from websites, automated filtering algorithms are applied. Web mining approaches can be evaluated based on the performance measures in information retrieval: the precision and the recall. Considering all relevant websites helps to improve the recall measure in information retrieval and considering results of a high quality filtering helps to improve the precision measure. Some web mining approaches are applied at successive points in time to enable an environmental internet tracing. They discover current trends and relevant changes from the internet (Loh, Mane, & Srivastava, 2011). Thus, internet information is a valuable source for strategic decision making in an organization.

2.2. Identification of signals and signal tracing

A well-known concept for implementing an early warning system used in strategic planning is introduced by Ansoff (1975) that focusses on the identification of signals, specifically weak signals. Signals are defined as events, e.g. future trends, changes, or further emerging phenomena with a specific impact on a given target (Yoon, 2012). It could be distinguish between strong signals and weak signals. A strong signal impacts a target at present above a specific threshold and it is expected that this signal also will impact the target in future (Mendonça, Cardoso, & Caraça, 2012). In contrast to this, a weak signal has none or a small impact on a target at present but possibly, it will get an impact on the target in future (Tabatabei, 2011). Thus, the identification of weak signals makes it possible for decision makers to be aware of events in advance that will impact the decision in future (Kuosa, 2010). A further definition of weak signals describes them as unstructured information with low content value at present time that reflects e.g. aspects of an opportunity or a threat without aiming at a specific target (Mendonça, Pina e Cunha, Kaivo-oja, & Ruff, 2004). If the content information becomes more concrete by mention the impact of the opportunity or threat on a specific target then a weak signal has become a strong signal (Holopainen & Toivonen, 2012).

In the internet, many webpages can be found where strong signals are mentioned. This is because their impact on a specific target is already known and they are widely discussed on several websites, new articles, and internet blogs. Thus, strong signal with impact on a specific target occur high frequently in the internet. In contrast to this, weak signals occur low frequently in the internet because they lack a current impact on a target and thus, they are not attractive for discussion and seldom mentioned on websites, new articles, and blogs. However, it might be that a small number of authors recognize the future impact of a weak signal and describe it in the internet. These few documents are among the large amount of information available in the internet. The identification of these documents and thus, the identification of weak signals in the internet is difficult and many practical approaches fail because of this information retrieval problem (Schwarz, 2005).

Literature introduces two approaches that specifically are built to identify weak signals within the large internet information. A knowledge structure based clustering approach is introduced by Tabatabei (2011) and a semantic clustering approach is introduced by Thorleuchter and Van den Poel (2013a). Both approaches use a document collection crawled from the internet at a specific point in time. However, they do not use time series.

Time series are defined as sequences of data chronologically arranged (Hamilton, 1994). Several methodologies exist for analyzing time series, e.g. the use of regression analysis for time series forecasting (Graff, Escalante, Cerda-Jacobo, & Gonzalez, 2013) and the use of pattern recognition for time series clustering (Rodpongpung, Niennattrakul, & Ratanamahatana, 2012). The methodologies are applied in several application fields, e.g. statistics, signal processing, and weather forecasting. The advantage of time series is that events can be traced over time and thus, event changes can be identified. This advantage may also be useful for tracing weak signals in the internet.

2.3. Semantic identification of weak signals

The use of semantic clustering for weak signal identification by Thorleuchter and Van den Poel (2013a) is motivated by the fact that information in the internet is written by several people and in several different writing styles. Different words are used to express the same event. Semantic approaches (e.g. LSI) are in contrast to knowledge structure based approaches. They consider term dependencies and use eigenvector techniques from algebra (Jiang, Berry, Donato, Ostrouchov, & Grady, 1999) to discover classes (semantic textual patterns) from a document collection. The semantic textual patterns contain terms that occur together in parts of the documents but also terms that might occur in the document parts. While people in the internet use different wordings for expressing the same event, considering the aspects of meaning (with a semantic approach) is more promising than focusing on the aspects of words (with a knowledge structure based approach).

Thorleuchter and Van den Poel (2013a) introduce a weak signal maximization approach based on semantic clustering. Documents are collected from the internet. LSI can be used to calculate k clusters from the documents. This is done several times, each time for a different value of k. For each k, a human expert analyses the clusters manually to identify weak signals. If k is too small then it is expected that the clusters represent rather strong signals than weak signals. If k is too large then it is expected that the clusters represent only parts of weak signals. In this case, one weak signal is represented by several clusters. The number of clusters with one-to-one correspondences to an identified weak signal is calculated for each k. This k is selected where the corresponding number is maximal.

LSI can be used for semantic clustering. Beside LSI, modern approaches with better performance than LSI also can be used for semantic clustering, e.g. 'Non-Negative Matrix Factorization' (Lee & Seung, 1999), 'Probabilistic Latent Semantic Indexing' (Hofmann, 1999), and 'Latent Dirichlet Allocation' (Blei, Ng, & Jordan, 2003; Ramirez, Brena, Magatti, & Stella, 2012). However, they are of higher complexity than LSI and thus, the use of LSI is more comprehensible for the reader. The aim of this paper it to show the feasibility of the proposed approach and thus, it is sufficient to use LSI instead of the other approaches.

3. Methodology

Fig. 1 depicts the proposed methodology. Based on a given hypothesis, it uses web and text mining to collect related textual information from the internet (see Section 2.1). The information is processed semantically by LSI and by a weak signal maximization approach (see Section 2.3). As a result, weak signal time series are identified (see Section 2.2) and compared to the given hypothesis.

The given hypothesis is a textual description of a strategic decision problem. It should contain the relevant terms and the relevant term co-occurrences for the problem. Further, it should be formulized in brief, clear, and comprehensible. To identify information specifically related to the decision problem, a data collection step is processed at successive points in time $(t_0, t_1, \ldots, t_n)$. This step searches the internet for textual information (webpages, blogs, etc.) based on a set of search queries. Each search query consists of a set of terms. The assignment of terms to search queries can be done automatically by composing the relevant terms and the relevant term co-occurrences in different variations. An automated assignment is normally of low quality and thus, we recommend using human experts for building search queries manually. For each point in time $(t_0, t_1, \ldots, t_n)$, the search queries are executed and the full text of all retrieved results (furthermore they are named documents) is crawled. Filtering methods from text mining are used to reduce the number or the size of documents in a preprocessing step (see Section 3.1). As a result, a document collection is created for each point in time.

For the document collection of t_0, a term-by-document matrix is created. It consists of an unmanageable high rank r. LSI is applied on the matrix to reduce its rank to k (see Section 3.2). This can be interpreted as the creation of k semantic textual patterns that represent the document collection semantically. Some of these k patterns also represent weak signals. However, patterns change based on the selection of k. Thus, the parameter k has to be selected carefully to identify an optimal value of k that results in the best performance concerning weak signal identification. This is done by applying a weak signal maximization procedure where LSI is applied for a various number of k (see Section 3.3) to compare its results.

After selecting k, a latent semantic subspace for the document collection of t_0 is build. The impact of each document on each semantic textual pattern is calculated as well as the impact of each term on each semantic textual pattern. LSI separately projects the further document collections from t_1 to t_n into the semantic subspace. As a result, the impact changes of documents on patterns can be traced over time. This shows the increased, static, or decreased relevance of weak signals represented by semantic textual patterns.

3.1. Data collection and pre-processing

Data is collected with relevance to the given hypotheses. It is taken from the internet where a large amount of textual information can be found in websites, blogs etc. This information stems from different persons from different nations written in different languages and formulized in different writing styles. Thus, textual data in the internet is very inhomogeneous and thus, hard to process. However, this data source is well suited for identifying weak signals because it contains descriptions of new trends with future relevance to the given hypotheses.

Search queries executed by internet search engines are used to collect data. The search queries should cover all topics mentioned by the hypothesis. Creating these search queries is done manually in accordance to Thorleuchter and Van den Poel (2013b): relevant terms as well as relevant term co-occurrences are extracted from the hypotheses and composed to a set of search queries. The search queries are automatically executed one-by-one using the advanced programming interfaces of search engines. A set of documents is created based on the full texts of the retrieved search query results. The search queries are executed at different points in time and thus, several sets of documents are created.

Each set is pre-processed by discarding scripting code, images, and html-tags in the documents. Further, specific characters and punctuation are deleted. Tokenization is applied to identify single

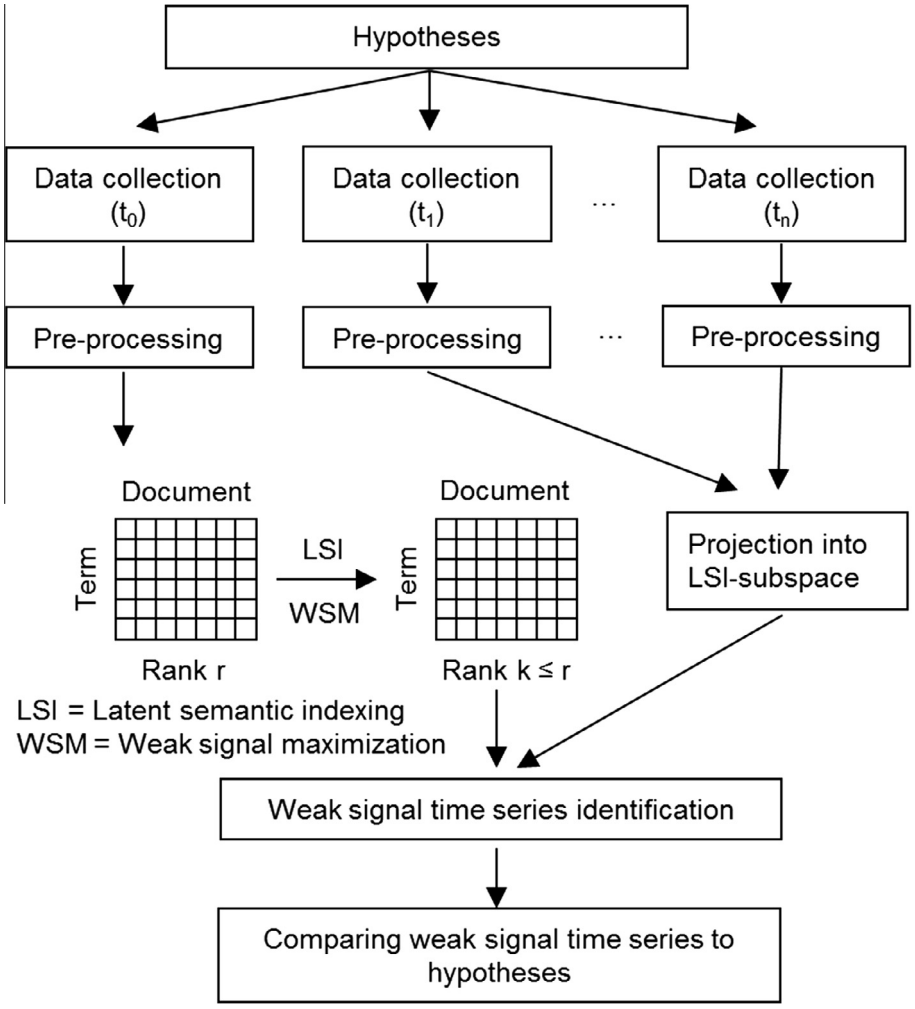

Fig. 1. Processing of the proposed methodology in different steps.

terms. Terms with typographical errors are corrected. Stemming is applied to summarize terms with the same stem. The information value of the summarized terms is estimated by stop word filtering and part-of-speech tagging. Low-informative terms are discarded. Further, terms are discarded if they occur only once or twice according to Zipf distribution (Zipf, 1949). Terms that occur in documents' paragraphs are compared to terms from the hypothesis to identify relationships between the paragraph and the hypothesis. This is done by applying Jaccard's coefficient as similarity measure. Paragraphs that are not related to the hypothesis are deleted within the documents. This leads to a reduced size of the documents (Thorleuchter & Van den Poel, 2013a).

For further processing, each document is transformed into a term vector in vector space model. The vectors' components are weighted frequencies instead of raw frequencies to improve performance (Prinzie & Van den Poel, 2006, 2007). A well-known weighting scheme is proposed by Salton, Allan, and Buckley (1994). It calculates the weight $w_{i,j}$ for term i and for a document j from a document set created at a specific point in time by

$$w_{i,j} = \frac{tf_{i,j} \cdot \log(n/df_i)}{\sqrt{\sum_{p=1}^{m} tf_{i,j_p}^2 \cdot (\log(n/df_{i_p}))^2}} \quad (1)$$

In this formula (1), n is the number of documents in the document set and df_i is the number of these documents that contain term i. Further, m is the number of distinguished terms in the document set. Formula (1) uses $\log(n/df_i)$ as the inverse document frequency and $tf_{i,j}$ as term frequency (Salton, Wong, & Yang, 1975). A length normalization factor can be found in the divisor to enable the comparison between documents of different lengths.

3.2. Latent semantic indexing

After data collection and pre-processing, a set of term vectors is obtained for each point in time $(t_0, t_1, \ldots, t_n)$. Term vectors of t_0 are used to build a term-by-document matrix A. The rank r of this matrix is large $(r \leqslant \min(m,n))$ because of the large number of documents in the document set and because of the large number of distinguished terms. This makes processing of the matrix nearly unmanageable. Reducing the rank can be done by applying LSI. It identifies semantic generalizations and it groups terms semantically in k clusters where k is smaller than r. The k clusters can be interpreted as semantic textual patterns that are latent in the data. The k clusters are used to build a new matrix A_k with rank k that is an approximation of A. In detail, LSI uses singular value decomposition to split the matrix A in three matrices

$$A = U \sum V^t \quad (2)$$

where Σ contains the singular values of matrix A on its diagonal components ordered by size $(\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_r)$. A weak signal maximization procedure is applied to identify an optimal value of k (see Section 3.3). Then, the singular values from k on $(\lambda_{k+1}, \ldots, \lambda_r)$ are discarded by creating a matrix Σ_k. Further, the columns of U and

V from $k+1$ on also are discarded by creating the matrices U_k and V_k. The approximation of A with lower rank k is calculated by

$$A \approx A_k = U_k \sum_k V_k^t \quad (3)$$

U_k and V_k define the LSI-subspace. U_k shows the impact of each term and V_k shows the impact of each document on each of the k semantic textual patterns. To project a new document d crawled from the internet at time t into the LSI-subspace, the corresponding term vector $v'_{d,t}$ of the new document is created by applying pre-processing steps (see Section 3.1). Deerwester, Dumais, Furnas, Landauer, and Harshman (1990) propose to transform the vector to a new vector $v_{d,t}$ that is comparable to the vectors of matrix V_k by

$$V_{d,t} = v'_{d,t} U_k \sum_k^{-1} \quad (4)$$

3.3. Weak signal maximization and times series identification

Reducing the rank of the matrix from r to k is normally done by machine based learning (Thorleuchter & Van den Poel, 2014a). For each k, a reduced LSI-subspace is built during training and the test examples are used to evaluate the value of k. This is done by applying logistic regression and n-fold cross validation to identify k with the largest area under the receiver operating characteristics curve (DeLong, DeLong, & Clarke-Pearson, 1988; Halpern, Albert, Krieger, Metz, & Maidment, 1996; Hanley & McNeil, 1982; Van Erkel & Pattynama, 1998). However, using a machine based learning approach for weak signal identification fails because weak signals occur low frequently and thus, the numbers of positive training and test examples are too small compared to the number of negative training and test examples.

To prevent this restriction, a weak signal maximization approach is proposed in literature (Thorleuchter & Van den Poel, 2013a). For each k, the k semantic textual patterns with impact above a specific threshold r are analyzed to identify weak signals standing behind the patterns. The number of one-to-one correspondences between semantic textual patterns and weak signals are calculated. This k is selected where the corresponding number of one-to-one correspondences is at its maximum (Thorleuchter & Van den Poel, 2014b).

As a result from weak signal maximization approach, semantic textual patterns are identified that represent weak signals. Human experts estimate the relationship between these semantic textual patterns and the given hypotheses. A pattern can contribute to the fulfillment of a hypothesis, it can be in contradiction to its fulfillment, or it can be indifferent.

The development of a weak signal can be shown by the number of documents with impact on the corresponding semantic textual pattern dependent on different points in time. While the number of all documents crawled at different points in time varies, the number should be calculated on a percentage basis. An increase of this value over time indicates that the weak signal (strongly) increases and probably will become a strong signal in future. A decrease shows that this signal probably will disappear in future. Otherwise, this also could show that a weak signal remains static.

Combining these weak signal time series to the relationship between the corresponding semantic textual patterns and the given hypotheses enables to show the overall impact of weak signals to the hypotheses.

4. Case study

A case study is applied to show the general feasibility of the methodology. We have used five different hypotheses and we have traced the signals at four different points in time. We are aware that the use of more hypotheses might improve quality of the results and we are aware that the use of time series at larger lengths might lead to a better understanding of the signals itself. However, using small values for hypotheses and points in time reduce complexity of the study. This increases comprehensibility and it enables the readers to get a better understanding about practical impacts of the proposed methodology.

The case study applies the proposed methodology in the field of storage technologies for renewable energy as an important aspect in energy transition. The main problem in energy transition is that production and consumption are often at different points in time. E.g. photovoltaic modules produce much energy during daytime and during summer season and little energy during the night and during winter season while consumption is often countercyclical to this. Energy storage could help to bridge this gap (Neupert et al., 2009).

Two of the most popular approaches are electrochemical and hydrogen based storages. Electrochemical batteries convert electrical energy into chemical energy by charging and vice versa by discharging. Chemical energy can be stored with good effectiveness of 90% and with high capacity but problems occur with self-drain and short live cycle. Hydrogen based storages use the energy to produce hydrogen from water. A reconversion to energy is done in fuel cells. This procedure has a large capacity and no self-drains however; the effectiveness is poor (about 40%) (Schiller, 2013).

Energy storages especially electrochemical and hydrogen based storages is an intensive field of research. The aim of this case study is to identify hypotheses from studies on the future of this field, to crawl related documents from the internet, to identify weak signals from the documents, to trace their development, and to evaluate their accordance to the developments described by these 'future studies'.

4.1. Hypotheses and data characteristics

Five hypotheses are manually extracted from 'future studies' (Evans, Strezov, & Evans, 2012; Löfken et al., 2013; Neupert et al., 2009; Schiller, 2013) in this field. They are depicted in Table 1 and they show developments of specific topics in the storage technology area of hydrogen and electrochemical components for energy transition.

Based on the given hypotheses, nine search queries are manually created in German language. Examples for these search queries translated in English language are presented below: "power to gas hydrogen energy transition", "energy storage renewable energy" and "electrochemical storage renewable energy". These search queries describe the storage technology area as defined by the extracted five hypotheses. They enable the identification of all internet documents dealing with this topic. These documents might confirm or disconfirm the extracted hypotheses.

The queries are automatically executed using Google API at four different points in time: in April 2013 (t_1), in July 2013 (t_2), in October 2013 (t_3), and in February 2014 (t_4). The analysis is restricted to websites in German language because Germany is leading in this technological field. Further, translation problems can be prevented

Table 1
Different hypotheses concerning the renewable energy sector.

1	Increased funding leads to advances in research and development
2	Increased usage of storages for intermittent energy sources from 2025
3	Power to Gas technology will get the key technology for long-term energy storage
4	Increased usage of lithium batteries in solar and automobile industry
5	Improved electrical grid stability by combining different storage types

by restricting to a specific language. In a further step, the search query results are crawled one-by-one where the textual information on the webpages is collected and stored in documents. A self-developed program is used for using Google API and for crawling.

A preprocessing step is applied (as described in Section 3.1). It uses the corresponding functions from SAS 9.3 Textminer to reduce the number and size of documents. As a result, the preprocessed documents represent the document collection of t_1, t_2, t_3 and t_4 respectively. The data characteristics are depicted in Table 2. Overall, it can be seen that this technological field is characterized by an increased number of relevant documents at subsequent points in time.

LSI and the proposed weak signal maximization approach are applied on the collected data by using SAS 9.3 Textminer. To identify an optimal value of k and of the threshold r, several rank-k-models are built. For each k, the maximal number of one-to-one correspondences is calculated based on different values of r. As a result, selecting $k = 25$ and $r = 0.4$ leads to six one-to-one correspondences as the overall largest number of weak signals (see Fig. 2). Thus, each of the corresponding six semantic textual patterns represents one and only one weak signal.

4.2. Results

The identified six weak signals are depicted in Table 3. For each signal, LSI calculates a list of semantically related terms (furthermore they are named signal-characteristic terms) with impact on the weak signal larger than threshold r. Examples for these signal-characteristic terms are shown in Table 3. These terms as well as the content of the corresponding documents are used to interpret the content of the corresponding weak signal manually.

To interpret the development of the identified weak signals, the percentage of related documents to a weak signal at a specific point in time is considered. For each weak signal and for each point in time, the number of documents from the document collection with impact on the corresponding semantic textual pattern greater than or equal to threshold r is calculated and divided by the number of all documents from the document collection. Table 4 shows the results for each point in time and interprets its development manually.

Each of the six identified weak signals impacts one or several hypotheses. An example is hypothesis two that is impacted by the weak signals one and four as shown below: The hypothesis forecasts an increased usage of storages for intermittent energy sources from 2025 (see Table 1). Weak signal one deals about improvement of energy storages based on advances in research and development and based on consumers' needs (see Table 3). Based on this information, a human expert notices that this weak signal contributes to the fulfillment of the hypothesis and the expert also sees that it increases strongly (see Table 4). Further, weak signal four describes that storages ensure to meet energy consumption needs of consumers in future (see Table 3). Thus, a human expert notices that this weak signal also contributes to the fulfillment of the hypothesis and it increases strongly, too (see Table 4). Overall, a human expert states that the hypothesis is in accordance to strong increasing weak signals one and four extracted from the internet (see Table 5).

A further example is that hypothesis three is impacted by three weak signals as shown below: The hypothesis predicts that the 'power to gas technology will get the key technology for long-term energy storage' (see Table 1). Weak signal three describes that hydrogen will become major energy carrier in future. It increases slowly (see Table 4). Weak signal six deals about integration of hydrogen/methane and lithium batteries as future storage concept. Weak signal two describes that power-to-gas technology will be used for long-long-term storage in the electrical grid (see Table 3). A human expert states that each of the three weak signals contributes to the fulfillment of the hypothesis. Weak signal three and six increase slowly and weak signal two remains static (see Table 4).

Overall, a human expert states that the hypothesis is in accordance to weak signals two, three, and six. However, time series show that the process of hypothesis fulfillment probably will get longer than expected (see Table 5).

As last example, the hypothesis five forecasts improved electrical grid stability by combining different storages types (see Table 1). Weak signal two says that Lithium battery will be used for short-term storage and power-to-gas technology will be used for long-long-term storage in the electrical grid. Weak signal six deals about integration of hydrogen/methane and lithium batteries as future storage concept (see Table 3). A human expert states that the hypothesis is in accordance to these weak signals. The weak signal two remains static and weak signal six increases slowly (see Table 4). This confirms hypothesis however; it the corresponding signal developments are too weak and thus, this hypothesis probably will not be fulfilled.

The impacts of weak signals on the hypotheses are depicted in Table 5 based on human expert's estimations. Weak signals without impact on a hypothesis are displayed in grey color. Weak signals that remain static do not support and not negate the fulfillment of a hypothesis thus; their impact is seen as indifferent. Slowly increasing weak signals only have little contribution to advance the hypotheses thus; their impact is seen as positive. Weak signals that increase strongly will become strong signals in future and thus, their impact on a hypothesis is strongly positive.

Table 5 is used to measure the precision and recall of the proposed approach. We are not aware about what will happen in future thus, the fulfillment of the hypotheses is seen as ground truth for this evaluation. This is because the corresponding 'future studies' are written by several human experts who extract the hypotheses manually from the large amount of internet documents. This can be seen as time-consuming work. We estimate the average time for creating a large 'future study' to at least two or three years. We set this baseline to 100% precision at 100% recall.

The strong increasing weak signals one, four, and five are in accordance to hypotheses one, two, and four. Thus, they successfully predict the hypotheses three times and the true positive (TP) is three. Hypotheses three and five are also in accordance to

Table 2
Data characteristics.

	t_0	t_1	t_2	t_3
Number of retrieved documents before pre-processing	1215	1373	1653	1728
Average results items per query before pre-processing	135	152	184	192
Number of retrieved documents after pre-processing	436	578	641	675
Average results items per query after pre-processing	48	64	71	75

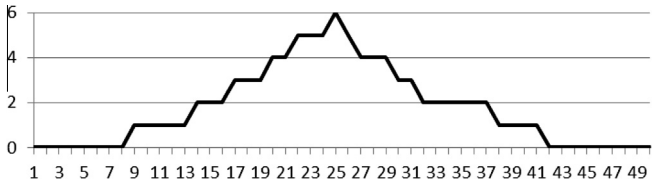

Fig. 2. Number of one-to-one correspondences (y-axis) based on the value of k (x-axis).

Table 3
Identified weak signals, signal-characteristic terms, and interpretation of signals' content.

Weak signal	Signal-characteristic terms	Interpretation of signal content
1	Energy, enhancement, storage, consumer, advance, development, increase, security of supply, ensure, environmental research,	Improvement of energy storages based on advances in research and development and based on consumers' needs.
2	Battery, short-term storage, long-term storage, methane, power-to-gas, distribution, grid, gas lain,	Lithium battery will be used for short-term storage and power-to-gas technology will be used for long-long-term storage in the electrical grid
3	Century, carrier, energy, supply, hydrogen, strengthening, growth, research, publication	Hydrogen will become major energy carrier in future
4	Transition, energy market, energy use, storing, independent, improve	Storages ensure to meet energy consumption needs of consumers in future.
5	Battery, lithium, types, sulfur, application, solar, building, car, sector	Application of different lithium batteries types in different sectors
6	Lithium-ion-batteries, hydrogen, project, methane, integrate, storage, solution, capacity	Integration of hydrogen/methane and lithium batteries as future storage concept.

Table 4
Percentage of related documents to a weak signal and its interpretation.

Percentage weak signal	t_0 (%)	t_1 (%)	t_2 (%)	t_3 (%)	Interpretation of signal development
1	17.28	28.93	35.69	37.31	Weak signal increases strongly
2	15.64	14.55	15.43	14.78	Weak signal remains static
3	16.46	18.20	20.45	21.63	Weak signal increases slowly
4	15.97	15.79	22.38	27.94	Weak signal increases strongly
5	14.65	25.19	28.37	31.05	Weak signal increases strongly
6	17.45	17.87	20.39	20.16	Weak signal increases slowly

Table 5
Impact of weak signals on the hypotheses (++ = strongly positive + = positive/0 = indifferent/grey colored cell = no impact).

Weak Signal / Hypothesis	1	2	3	4	5	6
1	++		+			
2	++			++		
3		0	+			+
4		0			++	+
5		0				+

weak signals however; these signals do not increase strongly and thus, the two hypotheses are not predicted as negative by the approach. This enables calculation the false negative (FN) to two and the false positive (FP) to zero. As a result, recall is calculated to 60% at 100% precision. Compared to the baseline (100% precision at 100% recall), it can be seen that the proposed semi-automated approach is of lower performance than a pure manual approach. However, the proposed approach saves times by analyzing the large amount of internet documents. Thus, the approach should be applied in cases where hypotheses about the future should be proposed in short time.

5. Conclusion

The new methodology – that is proposed in this paper – identifies weak signals from the internet to support strategic decision making. In contrast to previous work, it considers the time series of weak signals. The well-known LSI and weak signal maximization procedure are used to build a latent semantic subspace representing weak signals from an internet document collection. Internet documents are crawled at subsequent points in time to build further document collections. The documents are projected into the same semantic subspace. This enables to show the impact of the documents crawled at a specific point in time on each of the weak signals. Changes of these impacts over time can be traced to show the development of the identified weak signals. This could be the growth, the constancy or the disappearance of existing weak signals. Thus, the results support strategic decision makers by identifying trends and developments ahead of time. A case study is applied to show the general feasibility of the proposed approach. It identifies weak signal time series in the field of storage technologies for renewable energy. Their impact on hypotheses from studies on the future of this field is shown.

Future work should focus on aspects of data filtering. This is because the quality of semantic clustering results strongly depends on the quality of the given data. Existing approaches for an extended filtering of textual information should be considered to improve the performance of this approach. A further avenue of research is the use of PLSI, NMF, and LDA instead of using LSI. This might improve performance of the proposed approach. Last, the identification of weak signals from the LSI clusters is done manually up to now. Future work should focus on defining textual characteristics of weak signals. This helps to enable an semi-automated identification of weak signals by use of text mining approaches.

Acknowledgement

This work is processed using SAS 9.3 Textminer and a self-developed program for using Google web API and for webpage crawling.

References

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.

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

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.

DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics, 44*(3), 837–845.

D'Haen, J., Van den Poel, D., & Thorleuchter, D. (2013). Predicting customer profitability during acquisition: Finding the optimal combination of data source and data mining technique. *Expert Systems with Applications, 40*(6), 2007–2012.

Evans, A., Strezov, V., & Evans, T. J. (2012). Assessment of utility energy storage options for increased renewable energy penetration. *Renewable and Sustainable Energy Reviews, 16*(06), 4141–4147.

Graff, M., Escalante, H. J., Cerda-Jacobo, J., & Gonzalez, A. A. (2013). Models of performance of time series forecasters. *Neurocomputing, 122*, 375–385.

Halpern, E. J., Albert, M., Krieger, A. M., Metz, C. E., & Maidment, A. D. (1996). Comparison of receiver operating characteristic curves on the basis of optimal operating points. *Academic Radiology, 3*(3), 245–253.

Hamilton, J. D. (1994). *Time series analysis*. Princeton, New Jersey: Princeton University Press.

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology, 143*(1), 29–36.

Hofmann, T. (1999). Probabilistic latent semantic indexing. In *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.

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.

Kosala, R., & Blockeel, H. (2000). Web research: a survey. *ACM SIGKDD Explorations Newsletter, 2*(1).

Kuosa, T. (2010). Futures signals sense-making framework (FSSF): a start-up tool to analyse and categorise weak signals, wild cards, drivers, trends, and other types of information. *Futures, 42*(1), 42–48.

Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature, 401*(6755), 788–791.

Löfken, J.O., Lubbadeh. J., Honsel, G., Berkel, M., Hänssler, B., & Samulat, G. (2013). Stromnetze und Speicher. In *Review technology energie special* (pp. 64–88). Hannover: Heise Zeitschriften Verlag.

Loh, W. K., Mane, S., & Srivastava, J. (2011). Mining temporal patterns in popularity of web items. *Information Sciences, 181*(22), 5010–5028.

Mendonça, S., Cardoso, G., & Caraça, J. (2012). The strategic strength of weak signal analysis. *Futures, 44*(3), 218–228.

Mendonça, S., Pina e Cunha, M., Kaivo-oja, J., & Ruff, F. (2004). Wild cards, weak signals and organisational improvisation. *Futures, 36*(2), 201–218.

Neupert, U., Euting, T., Kretschmer, T., Notthoff, C., Ruhlig, K., & Weimert, B. (2009). *Energiespeicher*. Stuttgart: Fraunhofer IRB Verlag, pp. 9–63.

Prinzie, A., & Van den Poel, D. (2006). Investigating purchasing-sequence patterns for financial services using Markov, MTD and MTDg models. *European Journal of Operational Research, 170*(3), 710–734.

Prinzie, A., & Van den Poel, D. (2007). Predicting home-appliance acquisition sequences: Markov/Markov for discrimination and survival analysis for modeling sequential information in NPTB models. *Decision Support Systems, 44*(1), 28–45.

Ramirez, E. H., Brena, R. F., Magatti, D., & Stella, F. (2012). Topic model validation. *Neurocomputing, 76*(1), 125–133.

Rodpongpung, S., Niennattrakul, V., & Ratanamahatana, C. A. (2012). Selective subsequence time series clustering. *Knowledge-Based Systems, 35*, 361–368.

Salton, G., Allan, J., & Buckley, C. (1994). Automatic structuring and retrieval of large text files. *Communications of the ACM, 37*(2), 97–108.

Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM, 18*(11), 614–620.

Schiller, M. (2013). Hydrogen energy storage: the Holy Grail for renewable energy grid integration. *Fuel Cells Bulletin, 2013*(9), 12–15.

Schwarz, J. O. (2005). Pitfalls in implementing a strategic early warning system. *Future Studies, 7*(4), 22–31.

Tabatabei, N. (2011). *Detecting weak signals by internet-based environmental scanning* (Master thesis). Waterloo: Waterloo University.

Teo, T. S., & Choo, W. Y. (2001). Assessing the impact of using Internet for competitive intelligence. *Information & Management, 39*(1), 67–83.

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. (2014a). Quantitative cross impact analysis with latent semantic indexing. *Expert Systems with Applications, 41*(2), 406–411.

Thorleuchter, D., & Van den Poel, D. (2014b). Semantic compared cross impact analysis. *Expert Systems with Applications, 41*(7), 3477–3483.

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 Erkel, A. R., & Pattynama, P. M. T. (1998). Receiver operating characteristic (ROC) analysis: basic principles and applications in radiology. *European Journal of Radiology, 27*(2), 88–94.

Velásquez, J. D., Dujovne, L. E., & L'Huillier, G. (2011). Extracting significant websites key objects: a semantic web mining approach. *Engineering Applications of Artificial Intelligence, 24*(8), 1532–1541.

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.

Zipf, G. K. (1949). *Human behaviour and the principle of least effort*. Cambridge: Addison-Wesley.