Weak Signal Identification with Semantic Web Mining

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

[a] Fraunhofer INT, D-53879 Euskirchen, Appelsgarten 2, Germany, dirk.thorleuchter@int.fraunhofer.de

[b] Ghent University, Faculty of Economics and Business Administration, B-9000 Gent, Tweekerkenstraat 2, Belgium, dirk.vandenpoel@ugent.be URL: http://www.crm.UGent.be

[*] *Corresponding author at: Fraunhofer INT, Appelsgarten 2, 53879 Euskirchen, Germany. Tel.: +49 2251 18305; fax: +49 2251 18 38 305*

E-mail address: Dirk.Thorleuchter@int.fraunhofer.de (D. Thorleuchter).

Abstract

We investigate an automated identification of weak signals according to Ansoff to improve strategic planning and technological forecasting. Literature shows that weak signals can be found in the organization's environment and that they appear in different contexts. We use internet information to represent organization's environment and we select these websites that are related to a given hypothesis. In contrast to related research, a methodology is provided that uses latent semantic indexing (LSI) for the identification of weak signals. This improves existing knowledge based approaches because LSI considers the aspects of meaning and thus, it is able to identify similar textual patterns in different contexts. A new weak signal maximization approach is introduced that replaces the commonly used prediction modeling approach in LSI. It enables to calculate the largest number of relevant weak signals represented by singular value decomposition (SVD) dimensions. A case study identifies and analyses weak signals to predict trends in the field of on-site medical oxygen production. This supports the planning of research and development (R&D) for a medical oxygen supplier. As a result, it is shown that the proposed methodology enables organizations to identify weak signals from the internet for a given hypothesis. This helps strategic planners to react ahead of time.

Key Words: Weak Signal, Ansoff, Latent semantic indexing, SVD, Web Mining.

1 Introduction

A successful planning of research and development (R&D) requires an overview on current and future environmental conditions (Choi, Kim, & Park, 2007) to predict the arising of new technological approaches - the technology push - (Thorleuchter, 2008) and to predict changes in consumers' needs - the market pull - (Thorleuchter, Van den Poel, & Prinzie, 2010d) by time. Literature introduces a concept of environmental scanning (Abebe, Angriawan, & Tran, 2010; Tabatabei, 2011) that enables this prediction by extracting and analyzing information from the environment especially to identify events, trends, and relationships (Choo & Auster, 1993).

The concept of environmental scanning realizes a predictive view by applying a weak signal approach (Ansoff, 1975). A weak signal is an event or a development where an accurate estimation of its impact on a target (e.g. on organization's R&D) cannot be given because a single weak signal probably appears by chance (Ansoff, 1982). However, identifying several weak signals from different sources aiming at a common target is probably a hind that this target will be impacted in future. Thus, environmental changes can be predicted in advance that show future problem areas and opportunities. This enables the use of weak signals as early warning system for strategic planning.

As shown by Decker, Wagner, and Scholz (2005), the internet is a valuable information source for an environmental scanning and thus, for detecting weak signals. A website or a document itself is normally not a weak signal however; it might be that a website or a document contains a textual pattern that represents a weak signal (Uskali, 2005). Thus, a full text access to information in the internet is necessary to identify these weak signals. Performance reasons based on the large number of internet websites enforces a (semi-) automated approach e.g. web mining rather than a human based manual scanning (Gericke et al., 2009; Tabatabei, 2011).

Especially for R&D planning, information about three areas has to be considered (Thorleuchter, Van den Poel, & Prinzie, 2010c): the science for new technological aspects (technology push), the users for new product ideas (market pull), and the industry for new product development aspects (the link between technology and market). Technological

research results are described in articles published in scientific journals, in conference proceedings, and in various scientific document repositories. In recent years, access to the full text of these articles using the internet becomes much easier because of the increased number of open access journals and articles available today. Further, some publishers (e.g. Elsevier) offer open archives that enable a full text access to articles after a specific embargo period of time. Additionally, some publishers allow manuscript posting where accepted manuscripts can be posted on authors' personal or institutional websites. The Google Books initiative enables full text access to selected pages of conference proceedings published in books. This shows that in contrast to several years ago, the full text access to a large number of scientific articles is available today using the internet (Thorleuchter, Van den Poel, & Prinzie, 2010a).

Information about new product development can be found on companies' websites and in business magazines. Today, many magazines publish articles on their websites and thus, a full text access on this information is also available. Patents as representative for both, scientific results and industrial products are also published with full text in the internet (Thorleuchter, Van den Poel, & Prinzie, 2010b). Information about new product ideas from the users can be found in internet forums, blogs, micro blogs, review sites etc. The full text access to this information using the internet is possible, too. Overall, the planning of R&D can be supported by an environmental scanning and weak signals detection using the full text information in the internet today.

The proposed methodology uses semantic text classification combined with an automated web mining approach for environmental scanning and weak signals detection. This is in contrast to related research, where knowledge structure based text classification approaches are used (Yoon, 2012). The use of semantic text classification considers the fact, that weak signals are formulized by different persons, in different languages, and in different contexts. It might be that two textual patterns representing weak signals are related to a specific topic even if they do not share a common word. This relationship can only be identified with semantic approaches that consider aspects of meaning rather than aspects of words (Thorleuchter & Van den Poel, 2013b).

A further contrast to related research is the use of a new weak signal maximization approach. Existing literature that investigate latent semantic indexing as well known semantic approach apply prediction modeling approaches to calculate a performance optimized number of singular value decomposition (SVD) dimensions (Thorleuchter & Van den Poel, 2012e). They use training and test set that consists of a well-balanced number of positive and negative examples (Thorleuchter & Van den Poel, 2013a). The creation of a training and test set is not applicable to weak signal identification because weak signals for a specific topic occur low frequently. The number of positive examples for a specific topic is not sufficient to create a well-balanced training and test set. Further, an evaluation of weak signals' impacts can only be done considering the collection of all weak signals. Thus, a new weak signal maximization approach is proposed to identify the maximal number of weak signals for a specific topic to enable such an evaluation.

Up to now, the applied practical approaches for weak signal identification using a wide scope environmental scanning have failed. High tech companies in Europe had problems realizing a weak signal detection and evaluation because of the high manual effort caused by the lack of environmental scanning tools and the low quality of the results (Schwarz, 2005). Existing successful practical approaches for weak signal are restricted to a small number of documents e.g. 50 selected web pages (Decker et al., 2005) or financial news articles of one Finish newspaper (Uskali, 2005). Thus, the proposed semi-automated methodology bridges these gaps by implementing a web mining based environmental scanning and semantic weak signal identification. This enables a wide scope for environmental scanning, a low manual effort for human experts, and an improved identification performance.

In a case study, the proposed methodology is applied in the field of on-site medical oxygen production. R&D planners have provided a hypothesis concerning future developments. The methodology identifies relevant weak signals that are related to the given hypothesis. The weak signals do not verify or falsify the hypothesis; however they show that the hypothesis is in accordance to current trends extracted from the internet. This supports R&D planners by their decision making process.

Overall, a methodology is proposed that enables a practical use of the weak signal concept considering a wide scope of information from the internet, aspects of meaning, and

performance aspects to reduce the manual effort. Trends and developments can be identified in advance and they are a valuable source for R&D planners to support their decision making.

2 Background

2.1 Using Internet for R&D environmental scanning

The internet contains a huge amount of information and literature shows that the added value of this information outperforms the added value gained from using traditional information sources (D'Haen, Van den Poel, & Thorleuchter, 2012). Organizations use the internet in different ways e.g. for collecting and analyzing information from organization's customers (Alallak, 2010) and from competitive organizations (Teo & Choo, 2001) to advance organization's strategic planning (Purandre, 2008). Web mining approaches support organizations by information collecting because they offer an automated possibility to scan the internet for relevant information on websites (Kosala & Blockeel, 2000; Kobayashi & Takeda, 2000). They apply automated filtering algorithms to reduce the large number of websites identified by use of search engines. This is necessary to overcome performance restrictions because many retrieved and filtered results represent non-relevant information and thus, low precision values in information retrieval are obtained. Further, many relevant documents are not retrieved by the internet search engine. This leads to low recall values. In recent years, information about the R&D environment (science, industry, and consumer) is available and accessible in the internet as shown in the introduction chapter. This opened an opportunity to use the internet for R&D environmental scanning today.

2.2 Weak signals identification for R&D

The concept of weak signals has been introduced as early warning system to advance strategic planning (Ansoff, 1975; Tabatabei, 2011). It enables a timely identification of future events or developments that are relevant for a decision maker (Kuosa, 2010). Furthermore future events and developments are named topics. Literature introduces many different definitions of weak signals and most of them describe weak signals as unstructured

information with low content value (Mendonça et al., 2004). In a first stage, the weak signals reflect aspects of a threat or an opportunity. Then, their information content increases more and more e.g. they also describe the origin of a threat or an opportunity. Finally, weak signals become strong signals in a second stage and they indicate possible actions in future (Holopainen & Toivonen, 2012). Examples for weak or strong signals are articles in newspapers describing a specific topic, changes in sentiments of experts concerning this topic, and trends in the jurisdiction with impact on this topic (Mendonça, Cardoso, & Caraça, 2012). Strong signals point to a concrete topic that will occur with medium to high probability. A large number of strong signals for a specific topic can be found in the internet. This is because the topic is mentioned and discussed widely on several websites, in news articles, in internet blogs etc. Strong signals are not of interest for strategic planning because they occur too late for considering in strategic decision makings and thus, they do not provide a preview on environmental changes (Yoon, 2012). In contrast to the high frequent occurrence of strong signals concerning a specific topic, weak signals occur low frequently. Further, they can be used for strategic decision making because they occur early enough. For a specific topic, a small number of weak signals can be seen and it is hard to identify them from the large amount of information in the internet. This is the reason why many implementations of weak signal identification approaches fail in practice (Schwarz, 2005). Further, the occurrence of one weak signal is not sufficient for a predictive view, however; the occurrence of several weak signals that aim to the same topic might give a hind for future changes (Hiltunen, 2008). Thus, several weak signals have to be identified concerning the same topic and used for a strategic decision making (Ilmola & Kuusi, 2006; Rossel, 2009; Tabatabei, 2011).

Strategic R&D decisions are a subsection of strategic decisions. Descriptions of R&D topics are characterized by the occurrence of domain specific technical words (characteristic terms). This is in contrast to the colloquial language where the meaning of a specific term is often not clearly defined. Texts describing R&D topics also consist of an above-chance frequent co-occurrence of technical terms. This means that a specific technical term occurs more frequently together with a further term than it would be expected by chance. Both characteristics, the occurrence of characteristic terms and the frequent co-occurrences enable an easier identification of weak signals than it could be realized by using colloquial texts.

2.3 Latent semantic indexing for weak signals identification

With text classification, texts can be assigned to different classes. The classes have to be pre-defined in advance. This is done manually by human experts or automated by use of a set of training examples and machine based learning (Ko & Seo, 2009; Lin & Hong, 2011; Sudhamathy & Jothi Venkateswaran, 2012; Finzen, Kintz, & Kaufmann, 2012). Knowledge structure approaches are commonly used as instance-based learning algorithms for classification. Examples are the k nearest neighbor classification, simple probabilistic algorithms (e.g. naïve Bayes), decision tree models (e.g. C4.5), and support vector machine algorithms (Buckinx, Moons, Van den Poel, & Wets, 2004; Lee & Wang, 2012; Shi & Setchi, 2012). These approaches are already used for identification of weak signals theoretically (Tabatabei, 2011). However, they do not consider semantic aspects of the information. This is important because several weak signals for a specific topic have to be identified that are normally formulized by different persons. Thus, considering aspect of meaning is important to identify related weak signals. Further, some of the knowledge structure approaches do not consider dependencies of terms. Weak signals in R&D are characterized by the above-chance frequent co-occurrence of technical terms thus, it is also important to consider these dependencies.

In contrast to knowledge structure approaches, semantic approaches are better suited to consider aspect of meaning and to calculate term dependencies. The calculation of these semantic relationships between terms based on computational eigenvector techniques from algebra (Jiang, Berry, Donato, Ostrouchov, & Grady, 1999; Luo, Chen, & Xiong, 2011). Terms that occur together in a textual pattern are considered as well as terms that might be in this textual pattern (Thorleuchter & Van den Poel, 2012c; Thorleuchter & Van den Poel, 2012d). Semantic indexing is normally applied using LSI. Text patterns standing behind several documents from a document collection are identified (Park, Kim, Choi, & Kim, 2012). These text patterns also enable a clustering of the documents. They consist of a list of semantically related terms and the meaning expressed by the set of these terms is stated in different documents from the collection (Christidis, Mentzas, & Apostolou, 2012; Tsai, 2012). Further, the impact of each document on each text pattern is calculated. This considers well

term dependencies (Thorleuchter, Van den Poel, & Prinzie, 2012). Thus, LSI they can be used to identify weak signals.

Many modern approaches with better theoretical foundation and better performance than LSI have been introduced and applied in literature e.g. ‘Probabilistic Latent Semantic Indexing’ (Hofmann, 1999), ‘Latent Dirichlet Allocation’ (Blei, Ng, & Jordan, 2003; Ramirez, Brena, Magatti, & Stella, 2012), and ‘Non-Negative Matrix Factorization’ (Lee & Seung, 1999; Lee & Seung, 2001). In contrast to LSI, the improved approaches are of higher computational complexity than LSI. Applying them to a very large-scale document collection retrieved from the internet is difficult. Thus, LSI is used in the proposed approach to show the feasibility of this approach - in the knowledge that the performance can be improved by using the modern approaches instead of LSI. A very new and also very interesting approach is proposed by Ramírez and Brena (2012). Their query based topic modeling approach allows analyzing very large-scale collections with at least similar or even better performance than the above mentioned modern approaches by reducing computational complexity. As the case study (see Sect. 4) was already finished before we have been aware of this new approach, the use of this new approach might be interesting as an avenue of future research.

As mentioned in Sect. 2.2, weak signals occur with low frequency that means they occur on a few numbers of websites in the internet. Using weak signals as classes for text classification fails because the classes cannot be defined in advance by human experts. Further, a machine based learning approach cannot be performed because it could not be guaranteed that the number of positive training examples is above a specific threshold to ensure statistical significance for the classification results. This excludes the use of knowledge structure or semantic classification approaches. To overcome these limitations, the classes have to be defined by a semantic clustering approach where the identified semantic textual patterns are evaluated to identify weak and strong signals in a textual collection.

3 Methodology

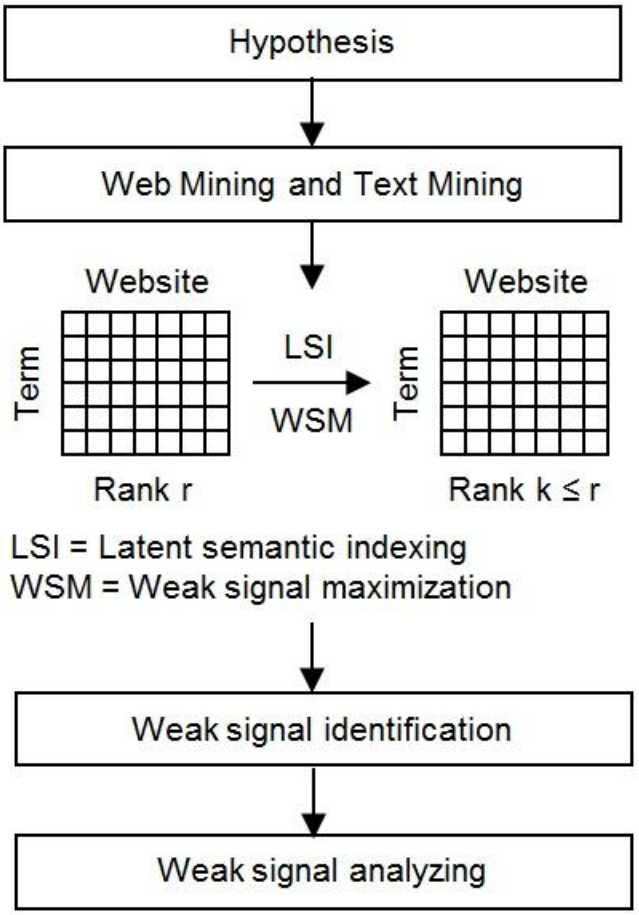

Fig. 1 shows the processing of the proposed methodology in different steps.

The proposed methodology in Fig. 1 identifies semantic textual patterns from the internet and it analyzes them to identify weak and strong signals. It applies an environmental search as described in Sect. 2.1 by using the internet. It considers the characteristics of weak signals as described in Sec. 2.2 by applying a semantic clustering approach (see Sect. 2.3).

The methodology starts with a hypothesis where an existing strategic decision problem is described. The weak signals are identified based on the topic described in this hypothesis. Thus, it is important that the hypothesis is formulized clearly and comprehensibly. The words that are used to formulize the hypothesis are considered for the next step, the creation of search queries. It is important that the search queries are created by human experts in high quality that many relevant documents are retrieved and that non-relevant documents are not

considered. The full text of the retrieved documents is crawled. Terms in the full text are compared to terms in the hypothesis to identify relevant sections within each document. These sections are used for further processing while the other sections are discarded. LSI is applied on the data and it creates a number of k different semantic textual patterns. They represent semantic aspects that occur on several different documents retrieved from the internet and thus, they can be used to represent strong and weak signals (see Sect. 2.2). The selection of k is based on a new weak signal maximization approach that leans on the weak signal characteristics from Sect. 2.2. This ensures that k is large enough to consider all weak signals and that k is small enough to discard semantic textual patterns that are not in accordance to the weak signal characteristics. The semantic textual patterns are split in weak signals and in strong signals. The relevance of each weak signal is analyzed manually concerning its impact on relevant terms. The development of a weak signal is calculated. As a result, the developments of weak signals are presented to the decision maker to improve strategic decision making.

3.1 Web Mining and Text Mining

Search queries are created manually to represent the hypothesis. A single search query is often not suited to cover a topic. Thus, several search queries have to be created to cover all different aspects of the hypothesis. To ensure an environmental scanning, it has to be considered that websites are written in different languages. Thus, created search queries for a specific language have to be translated in different languages. This should also be done manually by human experts to ensure a higher quality than a search engine can offer by automated translation.

The next steps are processed automatically. Each query is executed by an internet search engine and the search is restricted to the corresponding language from the query. The query results (the URLs of the websites) are used with a crawler that extracts the full text from all URLs. The retrieved results are stored as documents (one document per URL) in folders separated by the languages of the corresponding websites (Thorleuchter & Van den Poel, 2011b).

The full text from the retrieved documents are preprocessed and filtered to reduce complexity. In a first step, the raw test is cleaned from existing scripting code, images, and html-tags. Specific characters and punctuation are eliminated and typographical errors are corrected by using a dictionary from the corresponding language. Single words (terms) are identified by tokenization and case conversion is applied. In a second step, filtering methods are applied to reduce the number of terms for further processing. This includes stop word filtering (filtering of non-informative terms), via part-of-speech tagging (filtering specific syntactic category) up to stemming (reducing the number of terms with the same stem). Lemmatizing is not applied because existing practical methods are still error prone. The number of terms is reduced further by applying Zipf distribution (Zipf, 1949; Zeng et al., 2012).

The preprocessed full text from the retrieved documents consists of texts in different lengths from several bytes up to several megabytes that depend on the corresponding websites. The text normally is split in several sentences. However, a website could reflect several different topics. Then probably the text in one sentence is relevant for the topic described in the hypothesis and text in other sections or sentences is not. Tokenization is applied on the full text of each retrieved document and the term unit is used as a sentence. This means the text is split in its different sentences and a text similarity measure is used to compare the terms from each sentence to the terms from the description of the hypothesis. For this, term vectors in vector space model have to be build and Jaccard's coefficient can be used because it considers well the different lengths of term vectors (Thorleuchter & Van den Poel, 2011a). A similarity result value above a specific threshold shows that the corresponding sentence is relevant for the hypothesis and can be used for further processing. The other sentences are deleted. This reduces the lengths of the document and it ensures that the information used for latent semantic indexing is relevant.

The documents are written in different languages. However, the processing of latent semantic indexing requires that documents are written is the same language. The translation of documents to a target language can be done automatically e.g. by use of Google translate application programming interface (API). It offers an automated translation of a document collection. The quality of the automated translations is low compared to the quality from a manual translation of a human expert. However, a manual translation of each document

leads to high efforts because of the large amount of documents. Further, a high quality translation is not necessary because the translated text is transformed to term vectors in the next step. Thus, grammar aspects are not of interest and it is sufficient to translate relevant terms one-to-one. This can be done with automated approaches in good quality, too.

Term vectors in vector space model are created from the collection of all translated documents. For the components of the vectors, weighted frequencies are used. This is because literature shows that they outperform raw frequencies (Prinzie, & Van den Poel, 2007; Prinzie, & Van den Poel, 2006; Van den Poel, De Schamphelaere, & Wets, 2004). This is because weighted frequencies show the impact of a term on the collection of all documents (Sparck Jones, 1973). Large weights are assigned to terms that occur frequently in a very small number of documents and that do not or seldom occur in most of the documents from the collection (Salton & Buckley, 1988). The well-known terms weighting scheme proposed by Salton, Allan, & Buckley (1994) is used to calculate the weight $w_{i,j}$ for term i and for a document j 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 formula 1, the number of documents in the collection is n, the number of components from the term vectors is m, and the number of documents that contain term i is df_i. Further, inverse document frequency as represented by $\log(n/df_i)$ and term frequency ($tf_{i,j}$) are used (Chen, Chiu, & Chang, 2005). The divisor is a length normalization factor that considers different lengths of the documents.

3.2 Latent semantic indexing

The created vectors of weighted frequencies are composed to build a term-by-document matrix A with rank r ($r \leq \min(m,n)$). The large number of components from the term vectors leads to a large dimensionality of the matrix. Further, many terms only occur in a small number of documents and thus, the corresponding term vector component in many documents is set to zero. In total, this leads to term vectors with many zero values in the components and also to a term-by-document matrix that also consists of many zero values.

This fact is used by singular value decomposition to reduce the dimensionality of the matrix because it can be expected that the rank of the matrix is lower than its dimensionality. Singular value decomposition is a commonly used matrix factorization technique that is applied within LSI. A reduced matrix dimensionality leads to a summarizing of terms concerning aspects of meaning (Deerwester et al., 1990). Aspects of meaning are calculated based on the co-occurrences of terms in the documents. This enables to group semantically related terms together with high discriminatory power to other groups. The groups of terms represent semantic textual patterns (Thorleuchter & Van den Poel, 2012b). Singular value decomposition calculates singular values for each group and thus, for each semantic textual pattern. The singular values are sorted in descending order in a diagonal (r x r) matrix Σ. Singular value decomposition calculates two further matrices, too. The (m x r) matrix U shows the impact of terms on the semantic textual patterns and the (n x r) matrix V shows the impact of the documents on the semantic textual patterns. The calculation is shown in formula 2.

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

Matrix U and matrix V are used for calculating the weak signal maximization approach as introduced in Sect. 3.3.

3.3 Weak signal maximization approach for latent semantic indexing

Literature shows how to reduce the dimensionality of the matrix from r to k with singular value decomposition (Chen, Chu, & Chen, 2010). The data is split in a test and training set both containing a specific percentage of positive examples (Thorleuchter, Herberz, & Van den Poel, 2012). The training set is used to build a LSI-subspace (Zhong & Li, 2010) and the test set is projected into this subspace to calculate the predictive performance. The performance on each rank-k model is measured using the area under the receiver operating characteristics curve, logistic regression, and n-fold cross validation (DeLong, DeLong, & Clarke-Pearson, 1988; Hanley & McNeil, 1982; Halpern et al., 1996; Migueis, Van den Poel,

Camanho, & Cunha, 2012; Van Erkel & Pattynama, 1998). As a result, an optimal value of k based on computational complexity and on predictive performance is selected.

Existing approaches from literature cannot be used to identify an optimized number of semantic textual patterns representing weak signals from the internet. This is because weak signals occur low frequently as shown in Sect. 2.2. Thus, the percentage of positive examples in a randomly selected training and test set is very low and to obtain significant results, the training and test set have to be very large. The sets contain unseen documents retrieved by an internet search. For training and testing, these documents have to be evaluated concerning the occurrence of weak signals by human experts manually. This causes an unmanageable high manual effort.

Our proposed weak signal maximization approach identifies the value of k where the number of weak signals represented by low frequently occurred semantic textual patterns with a strong relationship to the given hypothesis is maximized. Singular value decomposition calculates k semantic textual patterns from the collection of all retrieved internet documents. It is characteristic for this processing that a small number of k leads to a small number of semantic textual patterns that are impacted by a large number of documents. These patterns occur frequently in the collection of all documents and thus, they are not weak signals. Thus, using singular value decomposition with a very small number of k does not lead to the identification of any weak signals.

A very large number of k leads to a small number of patterns impacted by a large number of documents and to a very large number of patterns impacted by a very small number of documents. As shown in Sect. 2.2., weak signals should occur at least more than once or twice otherwise they probably occur by chance. To identify a weak signal, the number of documents with impact on the pattern should be above a specific threshold. Thus, a very large number of k also leads to the identification of none weak signals in total.

Weak signals are not only defined by the number of impacted documents however, they should be related to the given hypothesis, too. Thus, relevant terms in a semantic textual pattern as defined by a term impact on the pattern above a specific threshold are compared

to relevant terms from the given hypothesis by using text similarity measures. A similarity value above a specific threshold shows that a low frequent semantic textual pattern is also related to the given hypothesis and thus, it can be defined as weak signal.

Patterns that are impacted by a large number of documents also are impacted by the large number of terms from the different documents. Thus, comparing these terms to the terms from the given hypothesis normally leads to a low similarity value. Furthermore, patterns that are impacted by a very small number of documents e.g. one or two are only impacted by a very small number of terms. Here, a low similarity value is also obtained by comparing them to terms from the given hypothesis.

Thus, a maximal number of weak signals can be identified if k is not too small and not too large. Several rank-k models are created and the number of weak signals is calculated for each model. Comparing is done by use of a similarity measure e.g. Jaccard's coefficient that considers well different lengths of input vectors because the term vectors created from the hypothesis normally is from different size than the term vectors created from the semantic textual patterns.

4 Case Study

The case study applies the proposed methodology to support R&D planners of a company that offers medical oxygen to hospitals. This is normally done in two different ways: off- and on-site production. Off-site production means that the medical oxygen is produced by oxygen generators in the company, stored in high pressure gas cylinders, and transported to the hospitals. On-site production means that the company sells oxygen generators to hospitals and the oxygen is produced in the hospitals. The production of medical oxygen is based on two different methods that separate air in his components. The cryogenic air separation uses a low temperature rectification principle based on the fact that gases have different temperatures for changing aggregation states. Applying this method can produce a large quantity of oxygen with high oxygen purity of more than 99% that also can be used to obtain liquid oxygen. The non-cryogenic air separation uses the pressure swing adsorption (PSA) principle. It produces oxygen with a purity of about 93%. The R&D department of the

company is responsible for the technical improvement of the oxygen generators (for both methods and for both, on- and off-site production).

Medical oxygen for hospitals in Europe has to meet the requirements of the European Pharmacopoeia where the oxygen purity is an important point. In the past, a hospital management had to buy oxygen at a purity of at least 99% in Europe and in many non-European states. This excludes the use of pressure swing adsorption principle for on-site oxygen production. Since mid-2011, the European Pharmacopoeia has changed the requirements for European states. After transposing this to national law, it allows an oxygen purity of 93% for hospitals if the used oxygen generators are certificated according to ISO 13485:2003.

Based on this legislation amendment, the R&D planners have stated a hypothesis: The use of PSA for medical oxygen on-site production (93% purity) in Europe will increase in future. This increase will be equally distributed in European states. New companies especially from the domain of machinery and plant engineering will become suppliers by offering PSA oxygen generators for hospitals in future.

The aim of the case study is to identify weak signals that are in accordance or that are not in accordance to the hypothesis.

4.1 Web Mining and Text Mining

Based on the given hypothesis, ten search queries are created in English language. Examples are ‘Medical +oxygen +high +purity’, ‘Oxygen +pressure +swing +adsorption +PSA’, and ‘Ultra +high +purity +oxygen +on +site +generation’. They describe the area of high purity medical oxygen and they enable the identification of all internet documents dealing with this topic. The search queries are translated in different languages: German (GE), French (FR), Polish (PO), Czech (CZ), and Romanian (RO) to cover different regions in Europe.

The queries are executed automatically using Google API in mid-2012 one year after the European Pharmacopoeia has changed the percentage because the process of transposing

this to national law is time consuming. The hyperlinks of all query results are stored separately into the different languages. For each retrieved website, a crawler is used to extract the textual information and to store it in a document. As a result, 14.792 plain text documents are created automatically with a size of 213 megabytes in total. Text mining methods are applied as mentioned in Sect. 3.1. This identifies several non-relevant documents and it reduces documents' sizes of relevant documents. Overall, 8375 documents with a size of 40 megabytes are obtained.

4.2 LSI with weak signal maximization

Based on the data collected in Sect. 4.1, LSI is applied together with the proposed weak signal maximization approach. The thresholds of the approach are determined manually in a two-step process by human experts. Starting values are determined in a first step and the results from several rank-k models are evaluated in the second step. Based on this evaluation, some parameter values are adapted (first step) and several rank-k models based on the adapted values are created and evaluated again (second step). This two-step process repeats until the parameter values are optimized.

As a result, the threshold for the impact of a document on a textual pattern as depicted in matrix V from -1 to 1 is determined to 0.4. The percentage of documents with impact greater than or equal to this threshold on a textual pattern to the number of documents in total is also determined to identify weak and strong signals: A value of 0% to 2% is representative for very low frequently occurred semantic textual pattern that might be occurred by change and can be discarded. Weak signals are identified from 2% on to 6% and strong signals are identified from 6% on to 10%. More than 10% means that the content of a semantic textual pattern can be found on more than on every tenth website. This information is normally well-known and thus, not relevant for forecasting.

Matrix U shows the impact of a term on a semantic textual pattern from -1 to 1. The threshold for identifying relevant terms is determined to 0.4. Thus, a term vector for each semantic textual pattern is built on terms with an impact greater than or equal to 0.4. The term vectors are compared to the term vector from the hypothesis where the threshold for the result value is determined to 0.3 to identify patterns that are related to the hypothesis.

4.3 Results

An optimal value of k is identified with k=16. Thus, 16 semantic textual patterns are identified. Three patterns can be identified that are impacted by more than 10% of all documents. Eight patterns are identified where the percentage of the impacted documents is between 0% and 2%. Further, two strong signals and three weak signals have been identified. Comparing them to the hypothesis shows that the strong signals are not related to the hypothesis. This is because they deal about cryogenic air separation techniques and the transportation of high pressure gas cylinders.

Two of the three weak signals are related to the hypothesis. The first weak signal with 4.2% impacted documents describes various aspects of PSA based oxygen generators producing 93% purity. This is in accordance to the first sentence of the hypothesis.

The second weak signal based on a small number of impacted documents (2.3%) and it shows a new technological development that enables PSA based oxygen generators to produce medical oxygen with 99% purity in a multi-stage process. The increase of purity can be realized with low additional costs. Oxygen generators based on this technique can be used in Europe as well as in many non-European states and they are independent of future legislation amendments. This weak signal is not in accordance to the first sentence of the hypothesis because in future, it might be that 99% purity PSA generators are increasingly used for on-site medical oxygen production.

Table 1: Number of documents with impact on the first and second weak signal in different languages compared to number of documents in total

Weak signal		EN	CZ	FR	GE	PL	RO	∑
1	Number of documents	182	9	66	68	3	24	352
	Percentage	52%	2%	19%	19%	1%	7%	100%
2	Number of	118	3	47	13	1	11	193

	documents Percentage	61%	2%	24%	7%	0%	6%	100%
	Number of documents in total Percentage	3449 41%	514 6%	1133 14%	1583 19%	582 7%	1114 13%	8375 100%

The number of documents with impact on the first and second weak signal is sorted by language (see Table 1) e.g. 182 documents in English language (EN)) have an impact on the first weak signal. Further 52% of all documents with impact on the first weak signal are English language documents. They are compared to the distribution of all 8375 documents e.g. 41% of all documents are English language documents.

Table 1 shows that more English and French websites are related to the first weak signal than it would be expected (increase from 41% to 52% and from 14% to 19%). In contrast to this, Romanian websites do not mention the weak signal as often as it would be expected (decrease from 13% to 7%). Table 1 also shows that on Czech and Polish websites the 93% purity PSA medical oxygen on-site production is seldom mentioned related to the number of Czech and Polish websites in total. Further, it shows that information about on-site medical oxygen generators with 99% purity often can be found on English and French websites. This is not in accordance to the second sentence of the hypothesis because the topic is not equally distributed in European states.

The URLs of the documents with impact on the first weak signal are also evaluated concerning companies' websites. As a result, the following list of companies is identified: 'Linde, Air Liquide, Praxair, Messergroup, Pangas, Westfalen-ag, Basigas, DA-Energietechnik, Iga-gas, Airco-Druckluft, Cryotec, aircom24, Airtexx, IGS, Oxymat, Oxyplus, Oxair'. This list contains established medical oxygen suppliers as well as companies in the domain of machinery and plant engineering. This is in accordance to the third sentence of the hypothesis.

4.4 Evaluation

In the case study, the proposed approach identifies two weak signals each represented by a semantic textual pattern. The identification is based on the assignment of retrieved internet documents to the two corresponding semantic textual patterns by LSI / WSM. The performance of this assignment is evaluated based on precision and recall. For each of the two semantic textual patterns, the number of documents that are correctly assigned to the pattern is the true positive (TP). The number of documents incorrectly assigned to the pattern is the false positive (FP) and the number of documents that are incorrectly not assigned to the pattern (missing documents) is the false negative (FN). Precision is defined as TP / (TP + FP) and recall is defined as TP / (TP + FN). The evaluation is processed for documents in English and German language because the number of these documents is sufficient for a statistical evaluation.

Table 2: Precision and recall for the assignment of English and German documents to the semantic textual patterns standing behind the first and second weak signal

Language	Number of Documents	First / Second Weak signal	TP	FP	FN	Precision	Recall
EN	3449	First	153	29	69	84%	69%
EN	3449	Second	96	22	33	81%	74%
GE	1583	First	48	20	26	71%	65%
GE	1583	Second	9	4	3	69%	75%

Table 2 shows that the average value is 76% for precision at 70% recall. This outperforms the average value of the frequent baseline (3% precision at 70% recall). Further, the precision and recall values for German documents are smaller than that of English documents. An explanation for this is that the German documents are translated to the English language automatically. This reduces quality of the translated documents and thus, this also reduces the corresponding precision and recall values.

5 Conclusion

This work proposes a new methodology that enables the automated identification of weak signals for strategic forecasting. Weak signals are extracted from an organization's environment as represented by internet information. Based on a given hypothesis about the future, related websites are identified and textual information is extracted. LSI with a new weak signal maximization approach is applied on this textual information to identify weak signals. The identified weak signals describe trends and future developments and it is analyzed whether they are in accordance to the given hypothesis or not. The websites standing behind an identified weak signal can be analyzed concerning language aspects to identify the spacial distribution of this weak signal. Further the website address also can be used to identify relevant organizations or companies related to the weak signal. This enables strategic planners to identify new trends, the spatial distribution of these trends, and the corresponding players (e.g. competitive organizations) ahead of time.

Future work should focus on the optimization of the parameters by applying a parameter selection procedure. This enables an improved weak signal maximization approach. Further, the development of weak signals over time can be investigated by this methodology. For this, web mining has to retrieve the data at different points of time. Then, one probably could see that new weak signals occur, that existing weak signals disappear, or that existing weak signals become strong signals.

Literature

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.

Alallak, B. (2010). Evaluating the adoption and use of Internet-based marketing Information systems to improve marketing intelligence. *International Journal of Marketing Studies*, 2(2), 87-101.

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.

Buckinx, W., Moons, E., Van den Poel, D., & Wets, G.(2004). Customer-adapted coupon targeting using feature selection. *Expert Systems with Applications*, 26(4), 509-518.
Chen, M. C., Chiu, A. L., & Chang, H. H. (2005). Mining changes in customer behavior in retail marketing. *Expert System with Applications*, 28(4), 773-781.
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.
Choi, C., Kim, S., & Park, Y. (2007). A patent-based cross impact analysis for quantitative estimation of technological impact: the case of information and communication technology. *Technological Forecasting and Social Change*, 74(2007), 1296-1314.
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.
Christidis, K., Mentzas, G., & Apostolou, D. (2012). Using latent topics to enhance search and recommendation in Enterprise Social Software. *Expert Systems with Applications*, 39(10), 9297-9307.
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*, doi: 10.1016/j.eswa.2012.10.023.
Finzen, J., Kintz, M., & Kaufmann, S. (2012). Aggregating web–based ideation platforms. *International Journal of Technology Intelligence and Planning*, 8(1), 32-46.
Gericke, W., Thorleuchter, D., Weck, G., Reiländer, F., & Loß, D. (2009). Vertrauliche Verarbeitung staatlich eingestufter Information - die Informationstechnologie im Geheimschutz. *Informatik Spektrum,* 32(2), 102-109.
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.
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.
Hiltunen, E. (2008). The future sign and its three dimensions. *Futures*, 40(3), 247-260.
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.
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.
Ko, Y., & Seo, J. (2009). Text classification from unlabeled documents with bootstrapping and feature projection techniques. *Information Processing & Management*, 45(1), 70-83.
Kobayashi, M., & Takeda, K. (2000). Information retrieval on the web. *Association for Computing Machinery*, 32(2), 144.
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, C.H., & Wang S.H. (2012). An information fusion approach to integrate image annotation and text mining methods for geographic knowledge discovery. *Expert Systems with Applications*, 39(10), 8954-8967.
Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. Nature 401(6755), 788-791.
Lee, D. D., & Seung, H. S. (2001). Algorithms for Non-negative Matrix Factorization. In: Advances in Neural Information Processing Systems 13. Proceedings of the 2000 Conference. MIT Press. pp. 556-562.
Lin, M-H., & Hong, C-F. (2011). Opportunities for Crossing the Chasm between Early Adopters and the Early Majority through New Uses of Innovative Products. *The Review of Socionetwork Strategies*, 5(2), 27-42.
Luo, Q., Chen, E., & Xiong, H. (2011). A semantic term weighting scheme for text categorization. *Expert Systems with Applications*, 38(10), 12708-12716.
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.
Migueis, V. L., Van den Poel, D., Camanho, A. S., & Cunha, J.F. (2012). Modeling partial customer churn: On the value of first product-category purchase sequences. *Expert Systems with Applications*, 39(12), 11250-11256.
Park, D, H., Kim, H. K., Choi, I. Y., & Kim, J. K. (2012). A literature review and classification of recommender systems research. *Expert Systems with Applications*, 39(11), 10059-10072.

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.

Purandre, P. (2008). Web mining: A key to improve business on web. IADIS European Conference Data Mining, (pp. 155-159).

Ramírez, E. H., & Brena, R. F. (2012). Query Based Topic Modeling: An Information-Theoretic Framework for Semantic Analysis in Large-Scale Collections. In: *Quantitative Semantics and Soft Computing Methods for the Web: Perspectives and Applications* (pp. 69-95), Information Science Pub., USA.

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.

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., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management,* 24(5), 513–523.

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

Shi, L., & Setchi, R. (2012). User-oriented ontology-based clustering of stored memories. *Expert Systems with Applications*, 39(10), 9730-9742.

Sparck Jones, K. (1973). Index term weighting. *Information Storage and Retrieval*, 9(11), 619-633.

Sudhamathy, G. & Jothi Venkateswaran, C. (2012). Fuzzy Temporal Clustering Approach for E-Commerce Websites. *International Journal of Engineering and Technology*, 4(3), 119-132.

Tabatabei, N. (2011). Detecting Weak Signals by Internet-Based Environmental Scanning. Master Thesis, Waterloo University: Waterloo.

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

Tsai, H.H. (2012). Global data mining: An empirical study of current trends, future forecasts and technology diffusions. *Expert Systems with Applications*, 39(9), 8172-8181.

Thorleuchter, D. (2008). Finding Technological Ideas and Inventions with Text Mining and Technique Philosophy. In C. Preisach, H. Burkhardt, L. Schmidt-Thieme, & R. Decker (Eds.), *Data Analysis, Machine Learning, and Applications* (pp. 413-420). Berlin: Springer.

Thorleuchter, D., Van den Poel, D., & Prinzie, A. (2010a). Mining Ideas from Textual Information. *Expert Systems with Applications,* 37(10), 7182-7188.

Thorleuchter, D., Van den Poel, D., & Prinzie, A. (2010b). A compared R&D-based and patent-based cross impact analysis for identifying relationships between technologies. *Technological Forecasting and Social Change,* 77(7), 1037-1050.

Thorleuchter, D., Van den Poel, D., & Prinzie, A. (2010c). Mining innovative ideas to support new product research and development. In H. Locarek-Junge, & C. Weihs (Eds.), *Classification as a Tool for Research* (pp. 587-594). Berlin: Springer.

Thorleuchter, D., Van den Poel, D., & Prinzie, A. (2010d). Extracting consumers needs for new products - A web mining approach. In *Proceedings WKDD 2010* (p. 441.). Los Alamitos: IEEE Computer Society.

Thorleuchter, D., & Van den Poel, D. (2011a). Semantic Technology Classification - A Defence and Security Case Study. In Proc. *Uncertainty Reasoning and Knowledge Engineering* (pp. 36-39). New York: IEEE.

Thorleuchter, D., & Van den Poel, D. (2011b). Companies Website Optimising concerning Consumer's searching for new Products. In Proc. Uncertainty Reasoning and Knowledge Engineering (pp. 40-43). New York: IEEE.

Thorleuchter, D., & Van den Poel, D. (2011c). High Granular Multi-Level-Security Model for Improved Usability. In: System Science, Engineering Design and Manufacturing Informatization 1 (pp. 191-194). New York: IEEE.

Thorleuchter, D., Herberz, S., & Van den Poel, D. (2012). Mining Social Behavior Ideas of Przewalski Horses. *Lecture Notes in Electrical Engineering*, 121, 649-656.

Thorleuchter, D., Schulze, J., & Van den Poel, D. (2012). Improved Emergency Management by Loosely Coupled Logistic System. *Communications in Computer and Information Science*, 318, 5-8.

Thorleuchter, D., Van den Poel, D., & Prinzie, A. (2012). Analyzing existing customers' websites to improve the customer acquisition process as well as the profitability prediction in B-to-B marketing. *Expert Systems with Applications*, 39(3), 2597-2605.

Thorleuchter, D., & Van den Poel, D. (2012a). Extraction of Ideas from Microsystems Technology. *Advances in Intelligent and Soft Computing*, 168, 563-568.

Thorleuchter, D., & Van den Poel, D. (2012b). Predicting e-commerce company success by mining the text of its publicly-accessible website. *Expert Systems with Applications*, 39(17), 13026-13034.

Thorleuchter, D., & Van den Poel, D. (2012c). Using NMF for Analyzing War Logs. *Communications in Computer and Information Science*, 318, 73-76.

Thorleuchter, D., & Van den Poel, D. (2012d). Using Webcrawling of Publicly-Available Websites to Assess E-Commerce Relationships. In *SRII Global Conference* 2012 (pp. 402-410). San Jose, CA, USA: IEEE.

Thorleuchter, D., & Van den Poel, D. (2012e). Improved Multilevel Security with Latent Semantic Indexing. *Expert Systems with Applications*, 39(18), 13462-13471.

Thorleuchter, D., Weck, G., & Van den Poel, D. (2012a). Granular Deleting in Multi Level Security Models - an Electronic Engineering approach. *Lecture Notes in Electrical Engineering*, 1, 177, 609-614.

Thorleuchter, D., Weck, G., & Van den Poel, D. (2012b). Usability based Modeling for Advanced IT-Security - an Electronic Engineering approach. *Lecture Notes in Electrical Engineering*, 1, 177, 615-619.

Thorleuchter, D., & Van den Poel, D. (2013a). Technology classification with latent semantic indexing. *Expert Systems with Applications*, 40(5), 1786-1795.

Thorleuchter, D., & Van den Poel, D. (2013b). Protecting Research and Technology from Espionage. *Expert Systems with Applications*, doi: 10.1016/j.eswa.2012.12.051.

Uskali, T. (2005). Paying attention to weak signals: The key concept for innovation journalism. Innovation Journalism, 2(11), p. 19.

Van den Poel, D., De Schamphelaere, J., & Wets, G (2004). Direct and indirect effects of retail promotions on sales and profits in the do-it-yourself market. *Expert Systems with Applications*, 27(1), 53-62.

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.

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.

Zhong, J., & Li, X. (2010). Unified collaborative filtering model based on combination of latent features. *Expert Systems with Applications*, 37(8), 5666-5672.

Zipf, G. K. (1949). *Human Behaviour and the Principle of Least Effort.* Cambridge: Addison-Wesley.