Expert Systems with Applications 40 (2013) 4978–4985

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

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

Weak signal identification with semantic web mining

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

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

ARTICLE INFO

Keywords:
Weak Signal
Ansoff
Latent semantic indexing
SVD
Web mining

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.

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, 1984). 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, Thorleuchter, Weck, Reiländer, & Loß, 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

* Corresponding author. Tel./fax: +49 2251 18305.
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).

0957-4174/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2013.03.002

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, 2013d).

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, 2013). 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 (Kobayashi & Takeda, 2000; Kosala & Blockeel, 2000). They apply automated filtering algorithms to reduce the large number of websites identified by use of search engines (Thorleuchter & Van den Poel, 2013c). 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, Pina e Cunha, Kaivo-oja, & Ruff, 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