Technological Forecasting & Social Change 120 (2017) 59–76

Contents lists available at ScienceDirect

Technological Forecasting & Social Change

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

# Novelty-focused weak signal detection in futuristic data: Assessing the rarity and paradigm unrelatedness of signals

Jieun Kim[a], Changyong Lee[b,*]

[a] Institute for Data, Systems, and Society (IDSS), Massachusetts Institute of Technology (MIT), United States
[b] School of Management Engineering, Ulsan National Institute of Science and Technology, Republic of Korea

ARTICLE INFO

*Keywords:*
Weak signal
Novelty detection
Futuristic data
Text mining
Local outlier factor
Signal-portfolio map

ABSTRACT

Previous attempts to scan weak signals from quantitative data focus on earliness, but neglect the novel nature of signals. This study proposes an approach to novelty-focused weak signal detection from online futuristic data. For this, first, text mining is applied to extract signals in the form of keywords from futuristic data. Second, a local outlier factor is utilized to assess the rarity and paradigm unrelatedness of signals. The futuristic data is considered a source of weak signals and patent data is utilized as a proxy for existing paradigms of technological innovation. Finally, signal-portfolio maps are developed to identify the patterns of signal representations. The proposed approach helps broaden the source of weak signals and improve the sensitivity to the detection of weak signals. A case study on augmented reality technology is presented.

## 1. Introduction

In the current era of uncertainty in business, and a shortened lifecycle for many technologies, the significance of detecting weak signals has been a much highlighted topic in the field of technology foresight. Weak signals are defined as the early signs of future disruptions, discontinuities, trends, or other emerging big changes, that at first appear to be background noise but which have high potential to be a stronger pattern (Ansoff, 1975; Eckhoff et al., 2014; Saritas and Smith, 2011; Schoemaker et al., 2013). The typical examples of weak signals are messages and signs associated with early developments in technologies, societal innovations, conflicts, origins of conflicts, demographic shifts, new rivals, new regulations etc. (Day and Schoemaker, 2005; Saritas and Smith, 2011). Being able to scan and capture these signals serve important roles in predicting future change, in order to exploit new opportunities and avoid future threats (Rossel, 2009). Although Ansoff (1975) initiated the discussions on weak signals in a strategy literature, this study will focus on the technology-centered signs that are relevant to technology foresight. A significant body of work exists that has dealt with weak signals' conceptual and theoretical bases (Hiltunen, 2008a; Holopainen and Toivonen, 2012; Kaivo-oja, 2012; Mendonça et al., 2012; Rossel, 2012; Saul, 2006) and methods and models (Carbonell et al., 2015; Day and Schoemaker, 2005; Hiltunen, 2008b). The conventional method for weak signal detection has been to use environmental scanning and Ansoff's filters. The scanning systems, both conscious and unconscious, have three steps of filters: (1) a surveillance filter, including methods used in information acquisition; (2) a mentality filter, representing the selection that comes to the attention of a firm, and (3) a power filter, regarding the influence of managers who note or neglect information. They can be a useful dimension of knowledge management processes in organizations (Kaivo-oja, 2012).

Naturally, the usefulness of weak signal detection strongly depends on data sources and subsequent approaches. Many researchers have preferred human sources, such as futurists, scientists/researchers, colleagues, and consultants (Hiltunen, 2008a); however, the source has been expanding from human into textual and *online data* (Hiltunen, 2008a; Uskali, 2005), due to the considerable growth of the volume of publicly available data on the web (Decker et al., 2005; Thorleuchter and Van den Poel, 2015; Uskali, 2005). While scanning for weak signals in the past necessarily involved the challenges of manual gathering and analysis of massive online data, machine learning and data mining mitigates this limitation, facilitating a broad scan of the periphery where appropriate filters allow the identification of weak signals (Veugelers et al., 2010). In turn, they open the possibility for a more efficient analysis. Several recent attempts have been made to develop a technology-enhanced approach to detect weak signals from online sources (Eckhoff et al., 2014). Decker et al. (2005) suggested an approach to internet-based environmental scanning using information foraging theory, vector space models, and case-based reasoning. Tabatabaei (2011) uses a knowledge structure based approach and CLUTO software to cluster web data and identify weak signals.

* Corresponding author at: School of Management Engineering, Ulsan National Institute of Science and Technology, 50 UNIST-gil, Ulju-gun, Ulsan 44919, Republic of Korea.
*E-mail address:* changyong@unist.ac.kr (C. Lee).

http://dx.doi.org/10.1016/j.techfore.2017.04.006
Received 30 June 2016; Received in revised form 9 April 2017; Accepted 10 April 2017
Available online 20 April 2017

Schoemaker et al. (2013) developed a strategic radar which is an integrated framework that uses scenario planning, business analytics, and dashboard technology to monitor and scan for important external signals including weak signals. Yoon (2012) suggested the keyword-based weak signal detection approach in web news, quantitatively measuring the signal presence in media using degree of visibility, degree of diffusion, and their rates of increase; and identifies the signals that are relatively less visible and rapidly diffused as weak signals. Kim et al. (2013) suggested a framework in which the signal is reflected by the conceptual evaluation index and that the cross-impact estimation is achieved by a Bayesian network inference model. Thorleuchter and Van den Poel (2013, 2015) and Thorleuchter et al. (2014) use a latent semantic indexing approach to group documents with similar meaning for the further detection of weak signals, leveraging the information from the internet in a variety of forms, styles, and contexts.

While extensive, the previous research has had several shortcomings. First, data sources need to be selected more deliberately. Most of the papers mentioned, draw from the web indiscriminately (Decker et al., 2005; Schoemaker et al., 2013; Tabatabaei, 2011; Thorleuchter and Van den Poel, 2013, 2015; Thorleuchter et al., 2014) or merely one source, such as web news (Yoon, 2012). The former has a risk of including too much 'garbage' not related to the future of technology, while the latter runs a risk of biased information. In response to this dilemma, we suggest using "*futuristic data*", a future-oriented segment of online data from posts in weblogs, news, databases, discussion forums, Wikis, or social media content (Martinez and Walton, 2014), and "*patent data*". The use of futuristic data can filter irrelevant information while including a sufficient variety of contents from human and textual data or views of experts and the public, some of which are deeply novel and creative (Cachia et al., 2007; Pang, 2010; Schatzmann et al., 2013). Meanwhile, the patents serve as a major source for identifying existing technological trajectories and innovations (Abbas et al., 2014; Kim and Lee, 2015); since the individual patents are inventions and represent incremental progresses (Nomaler and Verspagen, 2016), they can be generally regarded as proxies for the *existing* paradigm of technological innovations, as opposed to hinting novel progress. By identifying portions of futuristic data that significantly depart from patent data, we can find the unrelatedness to the existing paradigms.

Second, in terms of the criteria for detecting weak signals, the existing research has not paid enough attention to the '*novelty*' of weak signals. Many studies have considered the strength or intensity of a sign in order to identify early but potentially impactful signs (i.e., *earliness*) as weak signals (Carbonell et al., 2015; Hiltunen, 2008b; Holopainen and Toivonen, 2012; Yoon, 2012), as in the first mention of global warming and climate change in the 1980s. However, one of the distinctive characteristics of weak signals is that they are anomalies or strange issues that are not compatible with the prevailing sense-making paradigm (Hiltunen, 2008b; Kim et al., 2010, 2013; Mendonça et al., 2012; Popper, 2010; Saul, 2006). They are "novel" in the true sense, embodying originality/rarity and paradigm modification/relatedness in creativity studies (Dean et al., 2006). While earliness is related to the degree of originality (Hasan et al., 2009), novelty is a more specific and distinctive concept of weak signal. Some early signals can be far from being novel; but novel signals are always early signals. In turn, we define weak signal in this paper as 'novel future signals (documents or keywords), in futuristic data that are not only *rare* but also have *paradigm unrelatedness*' (Dean et al., 2006). This conceptual definition is again interpreted as an operational definition of criteria used to assess signals in futuristic data.

Third, in terms of methodology, previous research has focused on crawling techniques for data collection; and text mining or semantic approaches for data processing. Consequently, data assessment and filtering in terms of effectiveness or relevance remains reliant on the manual efforts of the experts (Eckhoff et al., 2014). To achieve data-driven screening for novelty, we propose applying the *Local Outlier Factor* (*LOF*) (Breunig et al., 2000), a novelty detection technique, to futuristic data. It is a method used to identify 'new or unknown' data which a learning system is not aware of during training (Markou and Singh, 2003). Among other things, LOF has an advantage in that it can identify local outliers even if data is not uniformly distributed and has incoherent patterns. Since the futuristic and patent data are composed of diverse and fragmented contents, the LOF can easily treat this heterogeneity in analysis to identify novel weak signals that are detected as outliers. In addition, for data processing before assessing novelty, we utilize text mining (Berry and Kogan, 2010) to extract signals from a vast amount of information in futuristic and patent data and process documents into a structured keyword-document matrix. After assessing novelty, the results are visualized as a signal-portfolio map to explore and interpret the signals.

The remainder of this paper is organized as follows: Section 2 discusses the background of weak signal detection, futuristic data, and the LOF approach. Section 3 proposes the novelty-focused weak signal detection approach and illustrates the suggested approach of mining keywords, assessing novelty, and drawing signal-portfolio maps in AR technology. Section 4 discusses the results of our case analysis. The final section draws conclusions, highlights limitations and makes suggestions for future research.

## 2. Background

### 2.1. Futuristic data and patent data for weak signal detection

Until recently, the studies of weak signals have primarily drawn from expert-centered views. For example, Wygant and Markley (1988), who present the information life cycle of emerging issues, note that since an idea undergoes an 'elite awareness phase' before becoming accessible to the public for the first time, esoteric sources like artistic works, scientific fiction, the fringe and alternative press, and specialized journals are good for finding weak signals. However, with the emergence of ICT, recent studies have been compelled to extend the search for weak signals to sources with public-centered views (Cachia et al., 2007; Hiltunen, 2008a; Thorleuchter and Van den Poel, 2015). In technology foresight, ICT has surely improved foresight capabilities in terms of accessibility to future-relevant information, collaboration by the interactive participation of stakeholders, increased efficiency of foresight processes, linkages among different sources, market growth, advances in foresight methods, and quantitative data-handling (Keller and von der Gracht, 2014). That is, ICT has not only provided foresight and analysis tools, but has expanded information sources. Futuristic data, a collection of online extracted documents about the future of technology that incorporates a large participation by the experts and the public, is one such source of information that has been made available by the ICT (Kim et al., 2016a, 2016b; Kwon et al., 2017). In this paper, the "experts" are defined as the category of people who have expertise in future and/or technology; typical examples are futurists, researchers, scientists, engineers, and inventors.

Many recent studies have referred to the impact of futuristic data on technology foresight in terms of collective intelligence, creativity, and/or interdisciplinary expertise (Cachia et al., 2007; Haegeman et al., 2013; Schatzmann et al., 2013). They mentioned various concepts relevant to futuristic data but used different terminologies, such as futures 2.0 (Pang, 2010, 2011), foresight 2.0 (Nelson, 2010; Schatzmann et al., 2013), future oriented communities (Gheorghiu et al., 2009), collaborative foresight (Markmann et al., 2012; Weigand et al., 2014), online foresight platform (Raford, 2015), foresight support systems (Keller and von der Gracht, 2014; von der Gracht et al., 2015; Walden et al., 2000) etc. In line with these studies, this paper considers futuristic data as holding possible future depictions related to technology, generated by means of large-scale collaboration within virtual communities of experts and other ordinary people (Kim et al., 2016a).

As the source of weak signals, futuristic data have several advan-