Expert Systems with Applications 39 (2012) 12543–12550

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

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

Detecting weak signals for long-term business opportunities using text mining of Web news

Janghyeok Yoon

Department of Industrial Engineering, Konkuk University, 120 Neungdong-ro, Gwangjin-gu, Seoul 143-701, Republic of Korea

ARTICLE INFO

Keywords:
Weak signal
Future sign
Text mining
Web news
Peripheral vision
Business intelligence

ABSTRACT

In an uncertain business environment, competitive intelligence requires peripheral vision to scan and identify weak signals that can affect the future business environment. Weak signals are defined as imprecise and early indicators of impending important events or trends, which are considered key to formulating new potential business items. However, existing methods for discovering weak signals rely on the knowledge and expertise of experts, whose services are not widely available and tend to be costly. They may even provide different analysis results. Therefore, this paper presents a quantitative method that identifies weak signal topics by exploiting keyword-based text mining. The proposed method is illustrated using Web news articles related to solar cells. As a supportive tool for the expert-based approach, this method can be incorporated into long-term business planning processes to assist experts in identifying potential business items.

1. Introduction

In today's competitive business environment, the keyword 'future' is becoming more important because it can be directly connected with the identification of promising business opportunities for formulating long-term businesses (Yoo, Park, & Kim, 2009). Various methods for identifying future business opportunities range from customary approaches including brainstorming, voice-of-customer analysis and data envelopment analysis (Seol, Lee, & Kim, 2011) to specific approaches such as system evolution patterns (Mann, 2007; Yang & Chen, 2012; Yoon & Kim, 2011a), disruptive innovation theory (Christensen, 1997), weak signal analysis (Ilmola & Kuusi, 2006; Kerr, Mortara, Phaal, & Probert, 2006; Kuosa, 2010) and customized patent mining methods (Lee, Yoon, & Park, 2009; Yoon & Park, 2005; Yoon & Kim, 2011b, 2012).

Among these approaches, weak signal analysis has received much attention as a method for analyzing businesses of an uncertain future. In studies about the future, it has been concluded that futures cannot be forecasted by past inertia but are transformed discontinuously by interrupting events (Dator, 2002). Here, indications related to discontinuous transformations are generally called weak signals (Hiltunen, 2006). Weak signals are understood as advanced, noisy and socially situated indicators of change in trends and systems that constitute raw information material for enabling anticipatory action (Wikipedia., 2011). Weak signals including opinions and symptoms tend to be considered trivial, so they are usually recognized by pioneers or special groups, not by acknowledged experts (Hiltunen, 2008). Therefore, the peripheral vision to detect weak signals is important because it provides business experts with key concepts for identifying business opportunities of alternative futures.

Generally, the procedure for early warning scanning consists of four steps: (1) scanning weak signals, (2) assessing weak signals, (3) transforming the signals into issues, and (4) interpreting the issues for new futures (Maurits Butter et al., 2011). Among the steps, scanning weak signals is prerequisite for analyzing alternative futures. However, scanning weak signals has relied heavily on the intuitive insight of experienced-experts, whose services may be costly and not widely available and who may provide different results on weak signals. Furthermore, information sources including scientific articles, news and 7 blogs are now increasing exponentially in number and amount, so it is almost impossible to rely only on experts to scan weak signal topics for business intelligence.

In this regard, this paper presents a text mining procedure for scanning weak signal topics. The proposed quantitative procedure generates two types of keyword portfolio maps, the keyword emergence map and the keyword issue map, by using the occurrence information of keywords and time-weighted analysis. This automated method is expected to complement the expert-based approaches and can be incorporated into long-term business planning processes.

In Section 2, previous works on weak signals and text mining are overviewed. In Section 3, a procedure for scanning weak signal topics is proposed and is illustrated using Web news articles related to solar cells. Finally, Section 4 concludes the paper with future research topics.

E-mail address: janghyoon@konkuk.ac.kr

0957-4174/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2012.04.059

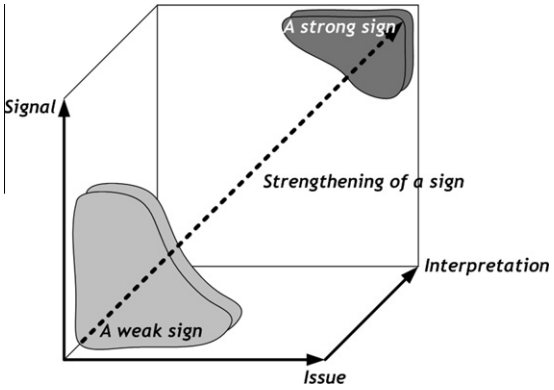

Fig. 1. Strengthening of the future sign, redrawn from Hiltunen (2008).

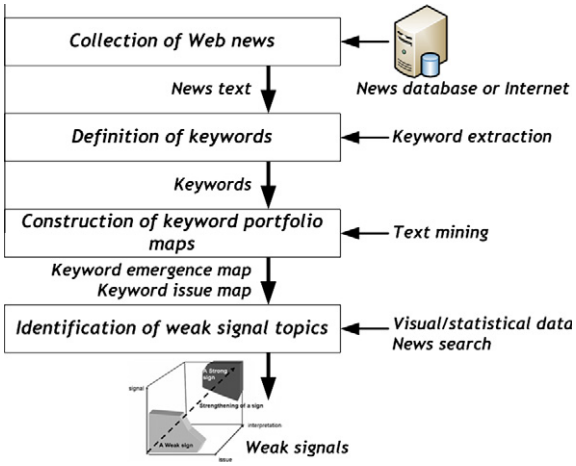

Fig. 2. A text mining-based procedure for peripheral vision.

2. Groundwork

The procedure proposed in this paper is based on weak signals and text mining, so previous works on this topic is overviewed in this section.

2.1. Weak signals

In studies about the future, weak signals are referred as the future-oriented information behind future trends, changes and emerging phenomena. Therefore, weak signals can act as indications that lead the discontinuity in trends and systems in the future, although they do not impact the present. Hiltunen (2008) defined 'future signs' as current oddities and strange issues that are thought to be key in anticipating future changes in different environments, and proposed three dimensions of future signs: 'signal' (the number and/or visibility of a future sign), 'issue' (the number of events that describe the diffusion of a future sign) and 'interpretation' (the receiver's understanding concerning the meaning of a future sign) (Fig. 1). In the three dimensional space, future signs strengthen from weak signs to strong signs.

Much research on the concepts of weak signals has been conducted to support strategic planning of corporations (Pirinen, 2010), to trigger employee's future-oriented thinking in analyzing a business environment (Hiltunen, 2007) and to analyze business environment and organizations (Day & Schoemaker, 2005; Ilmola & Kuusi, 2006; Rossel, 2009). Recently, many research groups including US Strategic Business Insight (SBI, 2011), UK Horizon Scanning Center (BIS, 2011) and Finland Futures Research Center (TrendWiki, 2011) developed processes and tools for monitoring weak signals, but their approaches are mainly based on the intrinsic knowledge of expert networks (Yoo et al., 2009).

For quantitative detection of weak signal topics, this paper adopts the three dimensional model of Hiltunen (2008) that conceptually describes the conditions of weak signals. Building on her model, this paper considers weak signals as emerging topics related to the keywords that are not much interpreted by people. For example, if the increasing rate of the occurrence frequency of a keyword is peculiar, then the keyword is strongly related to current oddities and strange issues. However, if the keyword has been rarely exposed to people, it is likely to be connected to weak signals. In this way, the method proposed in this paper identifies concepts that have a strong possibility of being weak signals, quantitatively and automatically.

2.2. Text mining

As a variation of data mining, text mining refers to the process of deriving high-quality information from text. This process is carried out as follows: (1) structuring the input text into structured data, (2) constructing analysis models, and (3) interpreting the output (Tseng, Lin, & Lin, 2007; Wang & Ohsawa, 2011). Text mining is different from regular data mining in that the patterns in text mining are obtained by processing natural language text rather than structured databases, by exploiting natural language processing and keyword matching.

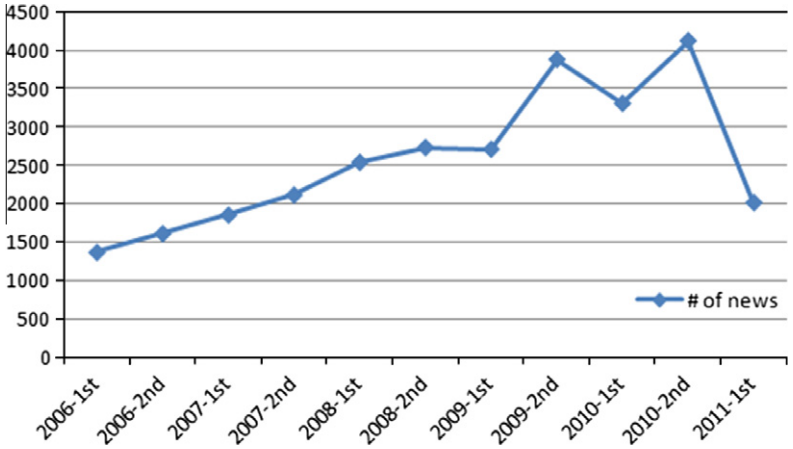

Fig. 3. The number of Web news articles.

Table 1
Selected keywords by layer.

Layer	Keywords
External factors [e]	Academy, regulation, army, environment, consortium, employment, environment, government, international, oil, political, pollution, tax, war, warm, water, ...
Business needs [n]	Capacity, attractive, awareness, clean, color, comfort, design, easy, entertainment, flexible, friendly, health, lifestyle, performance, portable, price, quality, reliable, safety, sustainable, ...
Product/technological characteristics [t]	Aircraft, building, appliance, chip, coal, combustion, conservation, dioxide, emission, fluorescent, hybrid, intelligence, interactive, large-scale, long-term, maintenance, medical, mobile, organic, panel, phone, plastic, plug-in, reactor, academy, recycling, stem-cell, vehicle, ...

In the keyword-based text mining, which is adopted in this paper, occurrence information including keyword occurrences and co-occurrence of keywords is widely used to construct analysis models (Park & Lee, 2011). In a given data, the occurrence frequency of a keyword is generally considered to be the measure of the importance of the keyword, although it implies some limitations (Salton & Buckley, 1988a, 1988b). Document frequency of a keyword is the number of documents in a collection in which the keyword occurs (Joho & Sanderson, 2007) and this is used to measure how widely a keyword is spread over a collection of textual information (Salton & Buckley, 1988a, 1988b). Co-occurrence of keywords may indicate interdependency or relationship among keywords, so relative importance of keywords can be statistically computed by using centrality measures of a network composed of keyword co-occurrences (Freeman, 1979; Lee & Jeong, 2008).

Building on the three dimensional model of future signs (Hiltunen, 2008), this paper exploits text mining techniques including the information techniques of occurrence frequency and document frequency to measure the visibility of future signs ('signal') and the diffusion of a phenomenon related to the future signs ('event').

3. Quantitative procedure for peripheral vision

The procedure proposed for text mining-based weak signal identification is composed of four steps (Fig. 2): (1) collecting news

Table 2
DoV of keywords by time period (a portion); $n = 11$, $tw = 0.05$.

Keywords	2006–1st	2006–2nd	2007–1st	2007–2nd	2008–1st	2008–2nd	2009–1st	2009–2nd	2010–1st	2010–2nd	2011–1st	Increasing rate
Environmental factors												
[e]academy	0.496	0.622	0.628	0.699	0.843	1.292	0.772	0.645	0.772	1.000	1.281	0.100
[e]regulation	1.917	2.246	2.541	2.780	2.756	3.772	3.635	5.143	5.500	5.190	5.221	0.105
[e]army	0.076	0.086	0.097	0.079	0.087	0.098	0.207	0.602	0.658	0.443	0.265	0.133
[e]environment	0.654	0.769	1.055	1.011	0.965	0.966	1.091	1.027	1.064	0.962	0.941	0.037
[e]consortium	0.018	0.016	0.010	0.014	0.015	0.011	0.022	0.030	0.026	0.028	0.030	0.052
[e]employment	0.028	0.022	0.041	0.025	0.037	0.051	0.078	0.097	0.110	0.086	0.072	0.101
[e]energy	2.661	2.452	3.101	2.876	3.053	3.560	4.369	4.083	3.953	4.454	4.353	0.050
[e]government	0.430	0.385	0.474	0.410	0.454	0.540	0.538	0.771	0.751	0.666	0.799	0.064
[e]international	0.149	0.198	0.192	0.207	0.218	0.428	0.246	0.285	0.290	0.314	0.319	0.079
[e]oil	0.790	0.550	0.561	0.505	0.774	0.860	0.490	0.494	0.548	0.456	0.562	−0.033
[e]political	0.097	0.101	0.090	0.105	0.093	0.122	0.082	0.114	0.089	0.085	0.102	0.005
[e]pollution	0.072	0.071	0.091	0.057	0.072	0.080	0.064	0.068	0.060	0.057	0.068	−0.006
...												
Business needs												
[n]attractive	0.067	0.080	0.085	0.082	0.075	0.094	0.076	0.091	0.093	0.096	0.102	0.042
[n]capacity	0.093	0.118	0.125	0.108	0.131	0.163	0.157	0.206	0.186	0.255	0.197	0.078
[n]price	0.780	0.791	0.835	0.857	1.008	1.156	0.993	1.064	0.956	0.950	0.915	0.016
[n]clean	0.738	0.588	0.745	0.712	0.701	0.900	0.816	0.870	0.883	0.847	1.086	0.039
[n]color	0.121	0.180	0.159	0.151	0.152	0.199	0.163	0.159	0.167	0.147	0.173	0.037
[n]comfort	0.042	0.050	0.058	0.050	0.039	0.041	0.063	0.068	0.052	0.050	0.042	−0.001
[n]compact	0.029	0.025	0.042	0.041	0.049	0.044	0.028	0.053	0.035	0.030	0.025	−0.014
[n]design	0.426	0.562	0.661	0.681	0.743	0.806	0.726	0.622	0.749	0.757	0.831	0.069
[n]entertainment	0.021	0.023	0.021	0.025	0.015	0.030	0.025	0.031	0.027	0.025	0.023	0.006
[n]flexible	0.024	0.031	0.035	0.039	0.037	0.053	0.030	0.038	0.051	0.047	0.053	0.082
[n]portable	0.021	0.016	0.022	0.026	0.036	0.047	0.023	0.021	0.042	0.033	0.027	0.026
[n]safety	0.066	0.065	0.089	0.120	0.088	0.189	0.141	0.101	0.121	0.101	0.136	0.076
...												
Product/technological components												
[t]appliance	0.061	0.065	0.089	0.071	0.061	0.073	0.075	0.090	0.061	0.052	0.075	0.021
[t]vehicle	2.850	3.046	3.231	3.690	3.487	4.388	4.042	4.181	4.233	3.936	4.482	0.046
[t]batteries	0.077	0.104	0.091	0.144	0.100	0.169	0.149	0.198	0.148	0.157	0.123	0.048
[t]building	1.247	1.417	2.041	1.613	1.635	1.719	2.031	2.111	1.964	1.864	1.801	0.037
[t]hybrid	0.162	0.110	0.144	0.173	0.137	0.184	0.147	0.159	0.147	0.097	0.106	−0.042
[t]large-scale	0.018	0.021	0.023	0.024	0.026	0.038	0.038	0.034	0.036	0.050	0.073	0.149
[t]long-term	0.051	0.053	0.055	0.064	0.068	0.063	0.076	0.069	0.077	0.073	0.094	0.064
[t]maintenance	0.037	0.032	0.039	0.028	0.051	0.050	0.077	0.067	0.099	0.090	0.089	0.093
[t]mobile	0.063	0.068	0.109	0.100	0.116	0.125	0.096	0.147	0.120	0.113	0.172	0.105
[t]organic	0.048	0.064	0.086	0.114	0.067	0.071	0.102	0.064	0.081	0.078	0.104	0.080
[t]plastic	0.042	0.062	0.081	0.082	0.066	0.084	0.078	0.099	0.093	0.083	0.073	0.058
[t]plug-in	0.026	0.034	0.019	0.035	0.039	0.057	0.046	0.026	0.029	0.018	0.022	−0.016
[t]remote	0.038	0.045	0.041	0.044	0.051	0.066	0.046	0.053	0.044	0.070	0.052	0.031
[t]sensor	0.031	0.072	0.041	0.069	0.063	0.116	0.044	0.058	0.051	0.070	0.086	0.106
[t]storage	0.055	0.068	0.063	0.102	0.077	0.150	0.095	0.112	0.113	0.146	0.142	0.100
[t]tree	0.213	0.298	0.319	0.342	0.373	0.359	0.356	0.335	0.341	0.377	0.352	0.052
...												

articles, (2) defining keywords concerning environmental factors, business needs and product/technological components, (3) constructing time-weighted keyword portfolio maps by using the occurrence information of the keywords, and (4) identifying weak signal topics by statistical analysis and news search.

3.1. Collection of Web news

Among the various types of information from technical documents including patents and journal articles to Web documents such as blogs, only Web news articles are used to identify weak signal topics in this paper because Web news is a more refined and reliable source of information than blogs or general Web pages. Furthermore, its contents cover a range of topics, from political, economic, social and business to technological. Therefore, Web news articles are appropriate for identifying potential weak signal topics for long-term business planning.

Although many methods and commercial databases are available for collecting Web news from the Internet (Reis, Golgher, Silva, & Laender, 2004; Zhang & Simoff, 2006; Zheng, Song, & Wen, 2007), the ProQuest database (http://search.proquest.com), which contains various information sources including newspapers, periodicals, dissertations and aggregated database of many types, was used to locate Web new articles in this paper. For analysis, a total of 28270 English Web news articles were located using keywords 'solar cells' and 'photovoltaic', for the period from Jan 1, 2006 to Mar 31, 2011. The number of news articles continuously increased, with the number of news articles increasing at a rate of about 13% every half year (period 2011–1st was excluded because it includes news articles only for 3 months). Each news article had an abstract, a full text and index fields including title, published data and publisher. In this step, all news articles were stored in an electronic format, such as html file and text file, for computerized preprocessing of textual information (Fig. 3).

3.2. Definition of keywords

Advanced keyword extraction methods using information filtering (Boger, Kuflik, Shoval, & Shapira, 2001), co-occurrence information (Matsuo & Ishizuka, 2004), domain knowledge (Hulth, Karlgren, Jonsson, Bostrom, & Asker, 2010) and graph analysis (Litvak & Last, 2008) are available for keyword suggestion. These extraction methods are commonly based on the frequency of a term in documents. Because the objective of this paper is not to suggest an effective way of extracting keywords, we simply identified a total of 140 keywords, which were extracted by solar energy experts after (1) removing irrelevant keywords by modifying English stopwords (Fox, 1989) and (2) identifying the occurrence frequencies of keywords. The keywords that are semantically similar but differently expressed were grouped into representative keywords; for example, 'automotive', 'bus', 'car' and 'motor' were converted into a representative keyword 'vehicle'. For identification of weak signal topics, the140 identified keywords were classified into representative layers including environmental factors, business needs and product/technological components. For example, keywords such as 'war', 'oil', and 'regulation' could be grouped into the environmental factor layer, keywords such as 'cheap', 'portable' and 'safety' into the business needs layer, and keywords such as 'batteries', 'vehicle' and 'hybrid' into the product/technological components layer (Table 1).

3.3. Construction of keyword portfolio maps

Building on the three dimensional model of Hiltunen (2008), this step constructs two types of keyword portfolio maps, the keyword emergence map and the keyword issue map. These maps help experts identify weak signal topics from the defined keywords and distinguish the weak signal topics from various future signs. According to Hiltunen (2008), weak signs are current oddities, so they are stranger than other future signs. Among Hiltunen's three

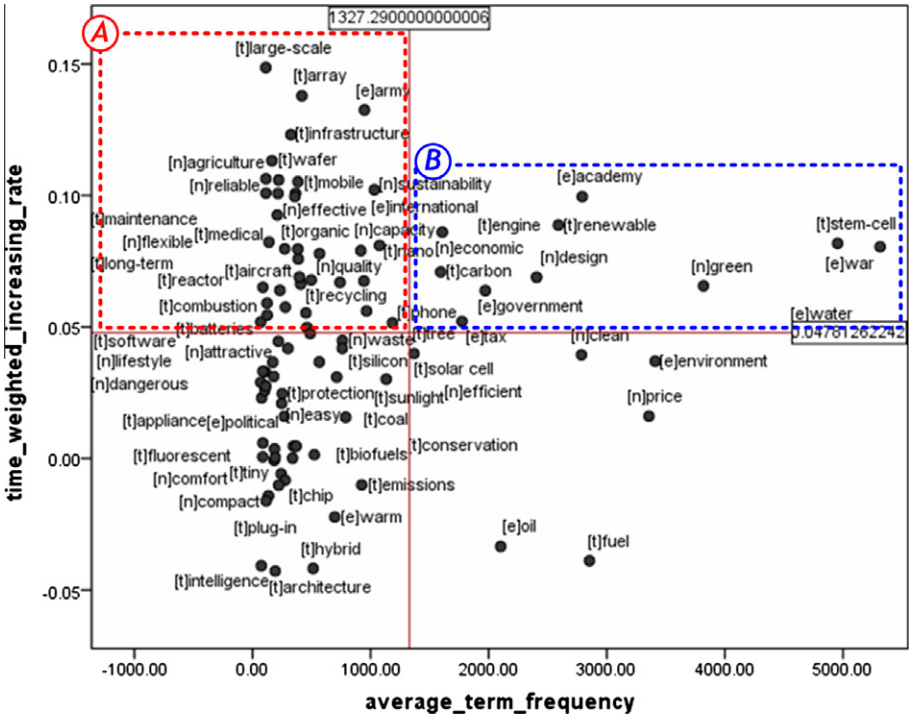

Fig. 4. Keyword emergence map (a portion); keywords in area A are connected with weak signals and keywords in area B are connected with strong signals.

dimensions of future sign, the axes 'signal' and 'issue' constitute the objective dimensions. Therefore, in this paper, these two levels of future signs are measured quantitatively by text mining of Web news articles. To this end, the two following general propositions are adopted (Jung, 2010).

Proposition 1. *Keywords of many occurrences in a collection are important.*

Proposition 2. *Recent appearances of keywords are more important than past appearances.*

This step exploits the occurrence information of keywords for proposition 1 and uses a time-weighted method to focus on recent oddities for proposition 2. Building on these propositions, two keyword portfolio maps, the keyword emergence map and the keyword issue map, are suggested for identification of weak signal topics.

3.3.1. Keyword emergence map

Among Hiltunen's three dimensions of a future sign, the 'signal' level of a future sign is related to the visibility of the future sign to the outside. With respect to text mining, this paper aims to measure the visibility of future signs using the occurrence frequencies of keywords. The frequency of a defined keyword in a set of news articles indicates the exposure degree of the keyword, so a high occurrence frequency of a keyword can be directly related to the high visibility of weak signal topics related to the keyword. Using a time weight, the degree of visibility (*DoV*) of keyword *i* in the period *j* can be defined as:

$$DoV_{ij} = \left(\frac{TF_{ij}}{NN_j}\right) \times \{1 - tw \times (n - j)\}, \quad (1)$$

where TF_{ij} is the total occurrence frequency of a keyword *i* in the period *j*, NN_j is the total number of news articles in the period *j*, *n* is the number of periods, and *tw* is a time-weight. *tw* increases as the environment around a given technology changes fast, while *tw* approaches 0 as the environment shows insignificant change. In this paper, *tw* was set to 0.05 after a careful review by three business experts on solar cells. Because the period 2011–1st does not include the news articles of six months, the proposed method computed the *DoV* of a keyword based on the occurrence frequency of the keyword per news article (Table 2).

Table 3
DoD of keywords by time period (a portion); $n = 11$, $tw = 0.05$.

Keywords	2006–1st	2006–2nd	2007–1st	2007–2nd	2008–1st	2008–2nd	2009–1st	2009–2nd	2010–1st	2010–2nd	2011–1st	Increasing rate
Environmental factors												
[e]academy	0.218	0.225	0.269	0.242	0.276	0.318	0.330	0.302	0.359	0.387	0.395	0.061
[e]regulation	0.422	0.470	0.504	0.545	0.573	0.626	0.678	0.701	0.748	0.787	0.837	0.071
[e]army	0.033	0.048	0.041	0.036	0.046	0.045	0.061	0.082	0.091	0.065	0.062	0.064
[e]environment	0.228	0.249	0.294	0.314	0.330	0.332	0.367	0.380	0.376	0.373	0.400	0.058
[e]consortium	0.011	0.013	0.008	0.009	0.013	0.008	0.012	0.017	0.016	0.018	0.016	0.035
[e]employment	0.016	0.015	0.021	0.016	0.021	0.029	0.048	0.050	0.057	0.053	0.042	0.103
[e]energy	0.370	0.382	0.446	0.465	0.512	0.564	0.612	0.639	0.669	0.710	0.730	0.070
[e]government	0.170	0.184	0.190	0.188	0.199	0.215	0.248	0.298	0.304	0.285	0.319	0.065
[e]international	0.094	0.119	0.121	0.122	0.135	0.156	0.156	0.165	0.175	0.174	0.174	0.063
[e]oil	0.213	0.180	0.202	0.201	0.242	0.254	0.202	0.230	0.211	0.196	0.209	−0.002
[e]political	0.059	0.053	0.058	0.063	0.052	0.064	0.055	0.065	0.056	0.048	0.065	0.010
[e]pollution	0.054	0.050	0.055	0.044	0.053	0.048	0.048	0.046	0.041	0.044	0.044	−0.020
...												
Business needs												
[n]attractive	0.050	0.058	0.064	0.056	0.060	0.064	0.061	0.068	0.069	0.072	0.079	0.047
[n]capacity	0.060	0.072	0.072	0.072	0.082	0.097	0.095	0.121	0.109	0.137	0.127	0.077
[n]price	0.265	0.282	0.298	0.308	0.346	0.365	0.366	0.390	0.375	0.361	0.375	0.035
[n]clean	0.232	0.240	0.252	0.259	0.272	0.305	0.310	0.315	0.326	0.321	0.340	0.039
[n]color	0.063	0.079	0.079	0.083	0.074	0.085	0.077	0.080	0.078	0.078	0.082	0.027
[n]comfort	0.034	0.041	0.042	0.040	0.032	0.035	0.042	0.043	0.041	0.043	0.034	0.001
[n]compact	0.020	0.018	0.030	0.030	0.034	0.035	0.024	0.024	0.027	0.021	0.021	0.008
[n]design	0.164	0.200	0.224	0.237	0.252	0.272	0.274	0.281	0.317	0.322	0.345	0.077
[n]entertainment	0.016	0.015	0.016	0.017	0.012	0.019	0.017	0.023	0.021	0.018	0.016	0.000
[n]flexible	0.021	0.023	0.022	0.029	0.028	0.034	0.022	0.029	0.034	0.033	0.032	0.043
[n]portable	0.015	0.011	0.016	0.019	0.023	0.026	0.018	0.017	0.027	0.022	0.022	0.036
[n]safety	0.044	0.046	0.059	0.069	0.056	0.066	0.061	0.066	0.077	0.067	0.082	0.065
...												
Product/technological components												
[t]appliance	0.043	0.037	0.052	0.048	0.039	0.047	0.046	0.056	0.043	0.039	0.045	0.003
[t]vehicle	0.412	0.455	0.497	0.536	0.569	0.611	0.641	0.682	0.720	0.756	0.795	0.068
[t]batteries	0.043	0.053	0.044	0.063	0.052	0.073	0.075	0.089	0.078	0.086	0.063	0.038
[t]building	0.299	0.330	0.379	0.375	0.385	0.429	0.461	0.503	0.498	0.515	0.535	0.060
[t]hybrid	0.059	0.049	0.058	0.068	0.055	0.075	0.064	0.064	0.059	0.045	0.055	−0.006
[t]large-scale	0.017	0.019	0.017	0.020	0.023	0.031	0.031	0.029	0.031	0.038	0.053	0.118
[t]long-term	0.051	0.053	0.055	0.064	0.068	0.063	0.076	0.069	0.077	0.073	0.094	0.064
[t]long-term	0.040	0.045	0.046	0.050	0.057	0.050	0.066	0.059	0.062	0.066	0.074	0.063
[t]maintenance	0.024	0.027	0.034	0.023	0.038	0.036	0.052	0.050	0.066	0.066	0.060	0.094
[t]organic	0.029	0.037	0.043	0.042	0.036	0.041	0.046	0.037	0.041	0.039	0.045	0.046
[t]plastic	0.028	0.037	0.038	0.052	0.045	0.046	0.044	0.052	0.051	0.043	0.038	0.033
[t]plug-in	0.011	0.009	0.010	0.022	0.020	0.026	0.024	0.016	0.017	0.012	0.014	0.028
[t]remote	0.030	0.033	0.028	0.033	0.035	0.038	0.032	0.037	0.029	0.043	0.037	0.021
[t]sensor	0.017	0.030	0.022	0.032	0.029	0.037	0.024	0.036	0.033	0.035	0.041	0.088
[t]storage	0.036	0.040	0.044	0.056	0.047	0.063	0.066	0.071	0.071	0.076	0.062	0.056
[t]tree	0.098	0.127	0.135	0.143	0.141	0.151	0.131	0.158	0.156	0.159	0.142	0.037
...												

According to Hiltunen (2008), future signs that have a possibility of being weak signals are the topics that have an abnormal pattern but are rarely exposed. Therefore, from the view of quantitative analysis, weak signal topics are likely to have related keywords of low absolute occurrence frequency but have a high range of fluctuation in the increase of occurrence frequency. Conversely, topics with keywords of low absolute occurrence frequency and a high increase rate of occurrence could be strong signals because they are considered to be important and exposed to the external. Therefore, the keyword emergence map can be obtained by using the average time-weighted increasing rate (geometric mean) and the absolute average term frequency of each keyword (Fig. 4).

3.3.2. Keyword issue map

'Issue' of Hiltunen's model is an axis that shows how much the phenomena of weak signal topics are disseminated over the outside. Appearance of a keyword in an article itself indicates the events of weak signal topics related to the keyword. Therefore, this paper measures the diffusion rates of phenomena related to weak signal topics using the document frequencies of keywords. In fact, document frequency is generally adopted to measure how general a term is in a collection of textual information (Salton & Buckley, 1988a, 1988b). Therefore, the document frequency of each keyword is directly related to issue level of a future sign. Using a time weight, the degree of diffusion (*DoD*) of keyword *i* in the period *j* is defined as:

$$DoD_{ij} = \left(\frac{DF_{ij}}{NN_j}\right) \times \{1 - tw \times (n - j)\}, \quad (2)$$

where DF_{ij} is the document frequency of keyword *i* in the period *j*, NN_j is the total number of news articles in the period *j*, *n* is the number of periods, and *tw* is a time-weight. Because the period 2011–1st does not include news articles of six months, the proposed method computed *DoD* of each keyword based on the document frequency of the keyword per news article (Table 3).

Similar to the determination of signal levels, future signs that have a possibility of being a weak signal are the topics that have an abnormal pattern but are rarely diffused. From the view of quantitative analysis, weak signal topics are likely to have low absolute document frequency but high range of fluctuation in the increase of document frequency. Conversely, topics with low absolute document frequency and high increase rate of document frequency could be strong signals because they are considered to be important and well spread. Therefore, the keyword issue map can be depicted by using the average time-weighted increasing rate of document frequency (geometrics mean) and the absolute average document frequency of each keyword (Fig. 5).

3.4. Identification of weak signal topics

Weak signal topics may have low signal and low issue levels, so this step identifies them by using time-weighted increasing rates and average term (document) frequencies of the related keywords; keywords related to weak signal topics are located in the area A of Figs. 4 and 5. This step identifies keywords whose signal (issue) levels are in the top 30% and of less than average absolute half-yearly term (document) frequency, and groups these keywords into two groups: weak signal keyword group and strong signal keyword group (Table 4). Using the keywords related to weak signal topics, this step searches the collected Web news articles and can explore potential weak signal topics from the text of the searched new articles (Table 5).

Commonly, the identified weak signal topics are considered promising but are rarely exposed to the external, so they can be used as valuable input for identifying potential business opportunities. Furthermore, unlike a previous work (Yoo et al., 2009), the proposed method can distinguish weak signal keywords from strong signal keywords, and the peripheral vision of the method can detect the recent phenomenon and oddities of future signs by exploiting a time-weighted method. Therefore, it is expected that the proposed method can be incorporated into scenario stud-

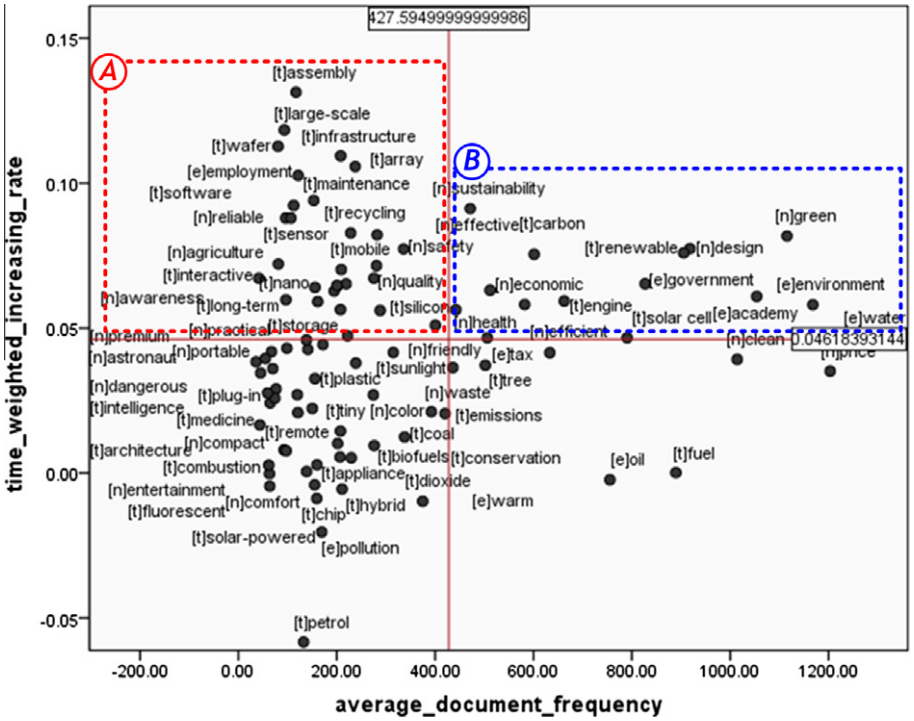

Fig. 5. Keyword issue map (a portion); keywords in area A are connected with weak signals and keywords in area B are connected with strong signals.

Table 4
Grouping of keywords related to future signs (top 30% in half-yearly time-weighted increasing rate).

Dimension	Keywords related to weak signal topics	Keywords related to strong signals
'Signal' (average term frequency = 1327)	[t]assembly, [t]large-scale, [t]array, [e]army, [t]infrastructure, [t]wafer, [n]agriculture, [t]sensor, [t]mobile, [n]sustainability, [n]reliable, [n]effective, [e]employment, [t]storage, [t]maintenance, [n]flexible, [n]economic, [t]organic, [t]medical, [e]international, [n]capacity, [n]safety, [t]aircraft, [n]performance, [n]health, [t]nano, [n]quality, [t]combustion, [t]long-term, [t]reactor, [t]plastic, [t]phone, [t]recycling, [n]awareness, [e]consortium, [t]tree, [n]friendly	[e]regulation, [e]academy, [t]renewable, [t]engine, [t]stem-cell, [e]war, [t]panel, [t]carbon, [n]design, [n]green, [e]government, [e]tax, [e]energy, [e]water
'Issue' (average document frequency = 428)	[t]assembly, [t]large-scale, [t]wafer, [t]infrastructure, [t]array, [e]employment, [t]maintenance, [t]software, [n]reliable, [t]sensor, [t]recycling, [n]effective, [n]capacity, [n]agriculture, [n]performance, [t]mobile, [n]quality, [t]interactive, [n]safety, [t]medical, [e]army, [t]aircraft, [e]war, [t]long-term, [n]awareness, [t]nano, [t]storage, [t]silicon, [n]health, [n]attractive	[n]sustainability, [n]green, [n]design, [t]renewable, [t]carbon, [t]stem-cell, [t]panel, [e]regulation, [e]energy, [t]vehicle, [e]government, [e]international, [e]academy, [t]building, [t]engine, [n]economic, [e]environment, [t]phone
Keywords related to future signs	[t]assembly, [t]large-scale, [t]array, [e]army, [t]infrastructure, [t]wafer, [n]agriculture, [t]sensor, [t]mobile, [n]reliable, [n]effective, [e]employment, [t]storage, [t]maintenance, [t]medical, [n]capacity, [n]safety, [t]aircraft, [n]performance, [n]health, [t]nano, [n]quality, [t]long-term, [t]recycling, [n]awareness	[e]regulation, [e]academy, [t]renewable, [t]engine, [t]stem-cell, [e]war, [t]panel, [t]carbon, [n]design, [n]green, [e]government, [e]tax, [e]energy, [e]water, [n]sustainability, [t]vehicle, [e]international, [t]building, [n]economic, [e]environment, [t]phone

Table 5
Identifying weak signal topics using selected keywords.

Keywords	Weak signal topics
[e]army	**army**'s solar tents, flexible solar panel for **military** solar power, portable solar chargers for **military** applications
[n]safety	health and **safety** concerns of photovoltaic solar panels, fears about the **safety** of nuclear power, solar panel electrical **safety**
[t]maintenance	low **maintenance** solar cell, protection and **maintenance** of solar panel, solar cell home **maintenance**
[t]nano	**nano**solar (a manufacturer of printable solar cells), super-thin solar panel from **nano**solar
[t]recycling	broken solar cells for **recycling** purpose, **recycling** the solar panels waste

ies and long-term business planning to detect weak signal topics in real-time, automatically.

4. Concluding remarks and future research

This research proposed a text mining-based approach to identify weak signals from Web news articles. Based on the three dimensional model of Hitunen's future signs, the proposed method: (1) collects news articles, (2) defines keywords concerning environmental factors, business needs and product/technological components, (3) constructs keyword portfolio maps by exploiting time-weighted analysis and occurrence and frequency information of the keywords, and (4) identifies weak signal topics by statistical analysis and news search. The proposed method was illustrated using Web news articles related to solar cells.

This paper demonstrated the possibility of quantifying the detection process of weak signals by text mining using Web news articles. The proposed method can detect weak signals more efficiently than human experts when dealing with the massive textual information due to the exponential increase in Web news articles in number and amount. Weak signals mostly tend to be recognized by pioneers or special groups but not by domain experts, so the proposed method can be incorporated into the business planning process to assist business experts in recognizing the driving factors of business environment changes and in developing potential business opportunities.

Despite these advantages, the proposed method has some challenges. First, the method was verified by using only Web news articles for weak signal detection. However, for real-time and automated analysis, future research should apply the proposed method to other Web information sources such as blogs, in addition to Web news articles, by exploiting the Web crawling techniques. This paper only considered Web news articles related to solar cells. To confirm the practicality of the proposed method, future research should investigate the topic of weak signals detections from Web news articles related to other technological fields.

Acknowledgement

This work was supported by the faculty research fund of Konkuk University in 2012.

References

BIS. (2011). Horizon Scanning Centre. <http://www.bis.gov.uk/foresight/our-work/horizon-scanning-centre>.
Boger, Z., Kuflik, T., Shoval, P., & Shapira, B. (2001). Automatic keyword identification by artificial neural networks compared to manual identification by users of filtering systems. *Information Processing & Management, 37*(2), 187–198.
Christensen, C. M. (1997). *The innovator's dilemma: when new technologies cause great firms to fail*. Harvard Business Press.
Dator, J. A. (2002). Advancing futures: Futures studies in higher education. Praeger Publishers.
Day, G. S., & Schoemaker, P. J. H. (2005). Scanning the periphery. *Harvard Business Review, 83*(11), 135–148.
Fox, C. (1989). *A stop list for general text*. New York, ACM: ACM SIGIR Forum.
Freeman, L. (1979). Centrality in social networks conceptual clarification. *Social networks, 1*(3), 215–239.
Hiltunen, E. (2006). Was it a wild card or just our blindness to gradual change. *Journal of Futures Studies, 11*(2), 61–74.
Hiltunen, E. (2007). The Futures Window-A Medium for Presenting Visual Weak Signals to Trigger Employees' Futures Thinking in Organizations. HSE Publications, working paper, w-423:.
Hiltunen, E. (2008). The future sign and its three dimensions. *Futures, 40*(3), 247–260.
Hulth, A., Karlgren, J., Jonsson, A., Bostrom, H., & Asker, L. (2010). Automatic keyword extraction using domain knowledge. *Computational Linguistics and Intelligent Text Processing*, 472–482.
Ilmola, L., & Kuusi, O. (2006). Filters of weak signals hinder foresight: Monitoring weak signals efficiently in corporate decision-making. *Futures, 38*(8), 908–924.
Joho, H., & Sanderson, M. (2007). *Document frequency and term specificity. Large Scale Semantic Access to Content (Text, Image, Video, and Sound)*. Pennsylvania: Pittsburgh.
Jung, K. (2010). *A study of foresight method based on textmining and complexity network anlaysis*. KISTEP: Seoul.
Kerr, C., Mortara, L., Phaal, R., & Probert, D. (2006). A conceptual model for technology intelligence. *International Journal of Technology Intelligence and Planning, 2*(1), 73–93.

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, B., & Jeong, Y. (2008). Mapping Korea's national R&D domain of robot technology by using the co-word analysis. *Scientometrics, 77*(1), 3–19.

Lee, S., Yoon, B., & Park, Y. (2009). An approach to discovering new technology opportunities: Keyword-based patent map approach. *Technovation, 29*(6–7), 481–497.

Litvak, M., & Last, M. (2008). Graph-based keyword extraction for single-document summarization. Proceedings of the Workshop on Multi-source Multilingual Information Extraction and Summarization, Stroudsburg, Association for Computational Linguistics.

Mann, D. (2007). *Hands-On Systematic Innovation for Business & Management*. IFR Press.

Matsuo, Y., & Ishizuka, M. (2004). Keyword extraction from a single document using word co-occurrence statistical information. *International Journal on Artificial Intelligence Tools, 13*(1), 157–170.

Maurits Butter, M. L., Cristiano Cagnin, Vicente Carabias, Totti Könnölä, Victor van Rij, Joachim Klerx, Petra Schape Rinkel, Effie Amanatidou, Ozcan Saritas, Jennifer Cassingena Harper, Lisa Pace. (2011). Scanning for early recognition of emerging issues; dealing with the unexpected, An operational framework for the identification and assessment of new future developments. Workshop Paper: SESTI Methodology, Workshop 26 October 2010.

Park, Y., & Lee, S. (2011). How to design and utilize online customer center to support new product concept generation. *Expert Systems with Applications, 38*(8), 10638–10647.

Pirinen, O. (2010). *Weak signal based foresight service*.

Reis, D. C., Golgher, P. B., Silva, A. S., & Laender, A. (2004). Automatic web news extraction using tree edit distance. *ACM*.

Rossel, P. (2009). Weak signals as a flexible framing space for enhanced management and decision-making. *Technology Analysis & Strategic Management, 21*(3), 307–320.

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

Salton, G., & Buckley, C. (1988b). Term-weighting approaches in automatic text retrieval∗ 1. *Information Processing & Management, 24*(5), 513–523.

SBI. (2011). The Scan™ Process. <http://www.strategicbusinessinsights.com/scan/process.shtml>.

Seol, H., Lee, S., & Kim, C. (2011). Identifying new business areas using patent information: A DEA and text mining approach. *Expert Systems with Applications, 38*(4), 2933–2941.

TrendWiki. (2011). TrendWiki Homepage. <http://www.trendwiki.fi/en/>.

Tseng, Y. H., Lin, C. J., & Lin, Y. I. (2007). Text mining techniques for patent analysis. *Information Processing & Management, 43*(5), 1216–1247.

Wang, H., & Ohsawa, Y. (2011). Innovation support system for creative product design based on chance discovery. *Expert Systems with Applications, 39*(5), 4890–4897.

Wikipedia. (2011). Futurology. <http://en.wikipedia.org/wiki/Futurology>.

Yang, C. J., & Chen, J. L. (2012). Forecasting the design of eco-products by integrating TRIZ evolution patterns with CBR and simple LCA methods. *Expert Systems with Applications, 39*(3), 2884–2892.

Yoo, S.-H., Park, H.-W., & Kim, K.-H. (2009). A study on exploring weak signals of technology innovation using informetrics. *Journal of Technology Innovation, 17*(2), 109–130.

Yoon, B., & Park, Y. (2005). A systematic approach for identifying technology opportunities: Keyword-based morphology analysis. *Technological Forecasting and Social Change, 72*(2), 145–160.

Yoon, J., & Kim, K. (2011a). An Automated Method for Identifying TRIZ Evolution Trends from Patents. *Expert Systems with Applications, 38*(12), 15540–15548.

Yoon, J., & Kim, K. (2011b). Identifying rapidly evolving technological trends for R&D planning using SAO-based semantic patent networks. *Scientometrics, 88*(1), 213–228.

Yoon, J., & Kim, K. (2012). TrendPerceptor: A property–function based technology intelligence system for identifying technology trends from patents. *Expert Systems with Applications, 39*(3), 2927–2938.

Zhang, D., & Simoff, S. (2006). *Informing the Curious Negotiator: Automatic news extraction from the Internet*. Springer.

Zheng, S., Song, R., & Wen, J. R. (2007). Template-independent news extraction based on visual consistency, Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999.