System Implementation for the Detection of Weak Signals of the Future in Heterogeneous Documents by Text Mining and Natural Language Processing Techniques

Israel Griol-Barres[1], Sergio Milla[2] and José Millet[3]
[1]*Instituto IDEAS, Vice-rectorate of Entrepreneurship and Employment, Universitat Politècnica de València (UPV), Camino de Vera s/n, 46022, Valencia, Spain*
[2]*FGYM, Vice-rectorate of Entrepreneurship and Employment, Universitat Politècnica de València (UPV), Camino de Vera s/n, 46022, Valencia, Spain*
[3]*Instituto ITACA, Universitat Politècnica de València (UPV), Camino de Vera s/n, 46022, Valencia, Spain*

Keywords: Weak Signal of the Future, Business Intelligence Architecture, Unstructured Information, Text Mining, Decision-Making.

Abstract: Not being able to cope with the constant changes in the market is currently one of the biggest threats for companies and start-ups. Therefore, the development of new systems to detect significant phenomena and future changes, is a key component for correct decision making that sets a correct course in the organisation. For this reason, a business intelligence architecture system is hereby proposed to allow the detection of discrete changes or weak signals in the present, indicative of more significant phenomena and transcendental changes in the future. In contrast to work currently available focusing on structured information sources, or at most with a single type of data source, the detection of these signals is here quantitatively based on heterogeneous and unstructured documents of various kinds (scientific journals, newspaper articles and social networks), to which text mining and natural language processing techniques (a multi-word expression analysis) are applied. The system has been tested to study the future of solar panels and the artificial intelligence sectors, obtaining promising results to help business experts in the recognition of new driving factors of their markets and the development of new opportunities.

1 INTRODUCTION

Business organizations from all around the world are affected from the fast pace of changes in the market, combined with their inability to manage and anticipate these changes (Brown and Eisenhardt, 1998). Markets are complex environments in which it can be very difficult to make the right decision and failing can mark the future of the organization.

The development of new processes that facilitate decision making in organizations considering data from different internal and external sources, thus becomes more and more important. Since the volume of data to be managed is increasing, these processes must involve techniques of capture, transformation, storage and automatic analysis that reduce the time and means required to analyse the data, while providing a high reliability.

Among them, data mining and business intelligent (BI) processes stands out (Khan, 2012). Business intelligence covers a broad category of applications and technologies to collect, analyse and provide access to huge amounts of data stored in a company’s database. With these data, models are created to interpret, predict and respond appropriately to the outside world (Fischler and Firschein, 1987). On the other hand, data mining is based on extracting relevant knowledge from different sources.

In the world of business competitiveness, "future" represents the need to identify potential new business opportunities (Yoo et alii, 2009), and many experts are working on evolutionary analysis, pattern detection, data mining methods, theories of disruptive innovation or detection of future signals of the future.

One type of future signal is the weak signal. This term was first coined in 1975 (Ansoff, 1975) as the detection of evidence of an emerging change within a continuous process of exploring a medium. That is, external or internal events and developments which

Griol-Barres, I., Milla, S. and Millet, J.
System Implementation for the Detection of Weak Signals of the Future in Heterogeneous Documents by Text Mining and Natural Language Processing Techniques.
DOI: 10.5220/0007380306310638
In *Proceedings of the 11th International Conference on Agents and Artificial Intelligence (ICAART 2019)*, pages 631-638
ISBN: 978-989-758-350-6

are still too incomplete to allow an accurate estimation of their impact and/or develop a response to them. However, these changes conceal the potential for more significant phenomena and transcendental changes in the environment, hence the importance of being able to identify and monitor them. If these phenomena evolve to become relevant (strong signals), they have the potential to reinforce a plan of action or to obstruct it. Another definition of the term weak signal (Godet, 1994) is that of "a change factor which is hardly discernible in the present but will constitute a strong tendency in the future."

Although in recent years "weak signals" have been receiving more attention in studies on the future, there is still no widespread use of the term, and there are authors who have used synonyms such as "seeds of change"(Molitor, 2003), "facts emerging" (Dator, 2005), "strategy signals"(Nikander, 2002) and "early warning signals" (Mannermaa, 1999).

Hiltunen (Hiltunen, 2008) introduced the concept of future sign which rapidly became the standard, which, at the same time, is based on the semiotic model of the sign (Peirce, 1868).

Based on the structure of this model, Hiltunen defined a triad on the future sign in these three dimensions: "issue", "signal" and "interpretation", with the aim of deepening the description of weak to strong signals (Kuusi and Hiltunen, 2007) and developed an analysis of the processes of significance in which the future signal is perceived, interpreted and produced. Figure 1 shows a comparison between both models.

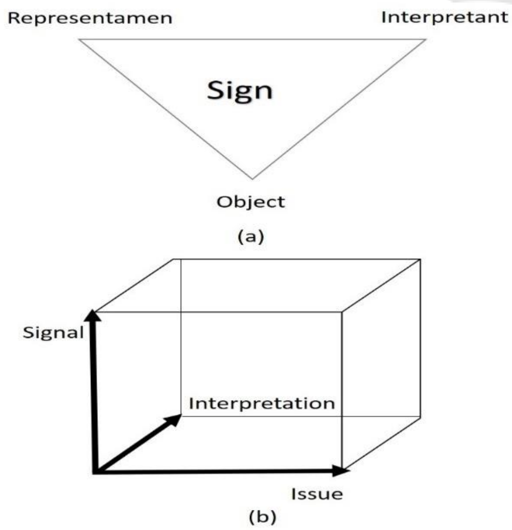

Figure 1: (a) Semiotic model of the sign by Peirce, and (b) Semiotic model of the future sign by Hiltunen.

Generally, a weak signal will have low absolute value in all three components, or at least in two of them. An example of the use of this model can be found in the Swedish clothing chain Hennes & Mauritz (H&M) which was selling old clothes at the price of new ones, under the label of "vintage". In fact, only 1% of its stores were carrying out this action. However, from the point of view of the future signal, the amount of signal (its visibility) was enormous thanks to successful marketing campaigns on newspapers. But the reality, the issue, is that 1% of the stores is not representative. This could be considered a weak signal since only one of the three components is strong, and therefore, it is not clear if this H&M action will represent a trend or not in the future.

Scanning for weak signals in the past involved gathering and analyzing massive online data manually. Now, the process is facilitated by machine learning and data mining techniques, thanks also to more efficient computational processing to detect weak signals from online sources (Eckhoff et aliis, 2014).

This article includes the design and implementation of a system to detect weak signals of the future. This system can be applied to any subject and it has been tested by the analysis of multiple unstructured documents concerning two sectors: (i) artificial intelligence and (ii) solar panels, with the goal of detecting words and expressions related to new future trends.

Section 2 of the paper describes related approaches and the state of the art; while section 3 covers the implementation of the system, and the results obtained through the experiment are explained in Section 4. Finally, the main conclusions of the study and future improvements are described in Section 5.

2 STATE OF THE ART

An organization operating in a complex and unpredictable environment must be flexible to detect this type of information. The implications of weak signals are very difficult to define at an early stage. However, in these complex environments, every organization is forced to make decisions in progressively earlier stages of knowledge. In addition, the time available to react is shortened, while the organization becomes more complex, making the detection of these weak signals of change a priority to make the right decisions. To do this, the organization needs to frequently scan and analyse the environment and this capacity of analysis will determine its ability to apply strategic behaviours.

Most of the publications about weak signals are related to a specific topic and a qualitative analysis. One instance is, the identification of weak signals and their sources related to terrorism and mass transport attacks (Koivisto et alii, 2016). The main problem for those systems is that can only be applied to that specific topic in specific conditions of time and space.

One of the first studies that includes a quantitative analysis (Yoon, 2012) only uses one type of source: web news. The quantification involved measuring the signal presence in media using degree of visibility and degree of diffusion (distinguishing number of documents and number of global appearances) and considering only their rates of increase with a time weight. Another work (Thorleuchter and Van den Poel, 2015) proposes a new methodology that enables the automated identification of weak signals for strategic forecasting.

To find weak signals encoded in a large number of documents, the process of future-oriented meaning must be taken into account (Jung, 2010): which is defined as the appearance and development of external themes and signals that are interconnected, their interpretation (the transfer of exosignals to endosignals), the re-creation of secondary exosignals for communication, and the different ways to act in consequence to these signals and the issues concerned.

In conclusion, detecting weak signals is a subtle process of observation and analysis, because the information regarding these signals is encoded in many data sources. For this reason, any system built for the detection of these signals will always contain processes that consume considerable memory and time resources and should therefore be optimized. Most of the studies up to now are qualitative, use a single source, are related to a specific topic or remain reliant on the efforts of experts.

This work describes a quantitative analysis of multiple sources and techniques in a system that can be applied to any topic, depending only on the documents used as its input.

3 IMPLEMENTATION OF A WEAK SIGNAL DETECTION SYSTEM

This section is divided into different parts. First, a system definition and a set of propositions will be established. Next, the methodology applied will be descripted and structured in six different phases of execution.

3.1 System Definition and Propositions

A system has been created considering the three variables of the Hiltunen semiotic model, including the interpretation variable which is related to the type of information source. The first step was to create annual internal repositories of information obtained from numerous online sources. These repositories were made up of scientific articles, journalistic articles and tweets related to a specific topic. Once the repositories were obtained, a system was required to manage a large volume of information from which to extract the necessary knowledge for decision making.

This study proposes the design of a system that quantitatively measures future signals using text mining techniques, that is, creating a tool that facilitates the analysis of an expert on a sector, assuming the following premises: (i) keywords with many occurrences in a collection of documents are important; (ii) recent occurrences of keywords are more relevant than past ones; (iii) more reliable results are obtain combining different data sources; (iv) false positives can be discarding by including natural language processing.

3.2 Description of the Methodology Applied

The proposed experiment is the design and implementation of a system focused on the search for weak signals related to artificial intelligence.

A key stage in this system is the Knowledge Discovery in Databases (KDD) methodology, which refers to the process of discovering knowledge and potentially useful information within the data contained in a certain information repository (Han and Kamber, 2001). When an organization uses a KDD system to discover potentially useful information, they usually start the process with their own databases of data from their activities, even though this info needs to be transformed in the KDD system. However, in this case a previous stage is necessary to generate a database with the collected information from several online sources: scientific journals, newspapers and social networks.

Figure 2 illustrates the steps of the system designed and implemented for our experiment. The process consists of 6 stages, which are (i) Collection and integration of information; (ii) ETL (Extract, transform and load); (iii) Selection, processing and transformation; (iv) Text mining focused on detecting weak signals; (v) Semantic Analysis; (vi) Interpretation and evaluation.

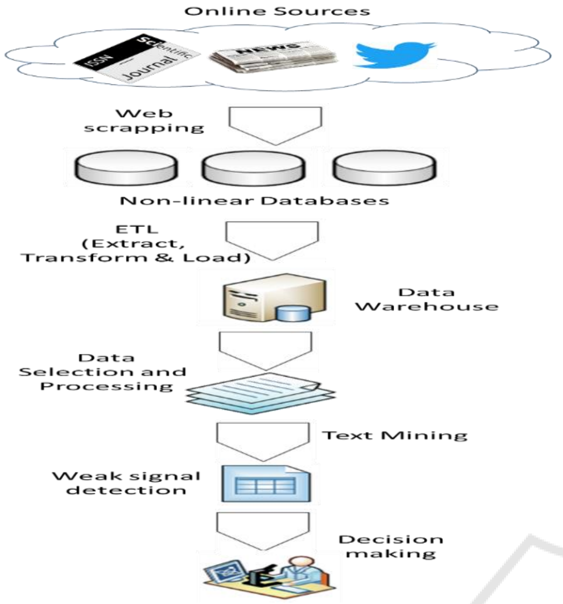

Figure 2: Proposed system for the detection of weak signals.

3.3 Phase 1: Collection and Integration of Information

For a company, knowledge can come from many different sources, such as information systems, reports, the internet, corporate databases, customers, suppliers or government agencies, or even the knowledge of employees. During this phase, a preliminary study of the sources from which the necessary information is obtained was carried out. These sources would be considered as technical and scientific types, journalistic sources and information stemming from social networks. Once the online sources were analysed, three were chosen: Direct Science as scientific source, the New York Times as journalistic source, and finally, Twitter as a social network. To collect the information, a Python algorithm was developed to extract data from HTML documents and tweets and to store it in NoSQL-type databases. The NoSQL documentary database chosen in this study is MongoDB, an open source document-oriented database.

Figure 3 shows a graph with the number of documents obtained per year from each of the three sources, related to the topic of artificial intelligence: 43.896 scientific papers extracted from ScienceDirect, 1.843 articles from the New York Times and 48.913 tweets.

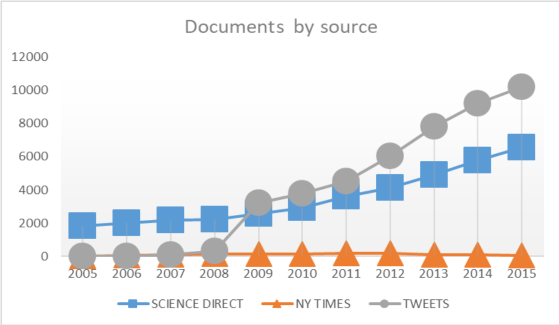

Figure 3: Number of collected documents per year and source related to the topic of artificial intelligence.

3.4 Phase 2: Extract, Transform and Load

Once the internal repositories are created, it is necessary to design and implement a data warehouse which supports the storage of a large volume of information. It is important to consider that this data warehouse must be issue-oriented and variable in time, and that the stored information cannot be lost (Immon, 2005w). At the same time an algorithm needs to be programmed to extract, transform and load the information in the warehouse.

The words extracted from the documents collected in the internal repositories of the information system are stored in the data warehouse created, as well as the following properties related to these words: document of origin, frequency of appearance in the document, year of appearance and source to which the document belongs.

This algorithm performs previous checks, to avoid inserting numbers, strange symbols and stop words, i.e., words that contribute neither to the semantics nor to the meaning of the text, and, therefore, can be dismissed as candidates for weak signals. Natural Language Toolkit (NLTK), a leading platform for building Python programs to work with human language data, was used for the elimination of stop words (Bird, Loper and Klein, 2009).

To reduce the time required to insert the words in the database, a stemming phase is performed at this point. The chosen word to represent this group of words is the one that appears most frequently in the creation phase of the database.

3.5 Phase 3: Selection, Processing and Transformation

To detect all possible weak signals, every word in the document apart from stop words and stems is considered a keyword for the experiment. The next step was to assign categories to each of these keywords. Categories were extracted from those provided in the scientific articles. In this way, every word is included in every category of the articles in which the word appears. A keyword can thus belong to more than one category.

3.6 Phase 4: Text Mining for Weak Signal Detection

Text mining is a variation of data mining applied to the process of obtaining high-quality information from a text (Ishikiriyama et alii, 2015). The process is characterized by structuring the input data, constructing analysis models and analysing the results obtained. The big difference with respect to data mining is that patterns in text mining are obtained by processing natural language rather than by processing structured databases.

The "signal" dimension of a future sign is related to the visibility of the future signal. To study the visibility of a sign, the absolute frequency of occurrence of each word in a set of documents (from the three different sources and for every year) is measured. This is defined as the Degree of Visibility (DoV) of the keyword i in period j, as can be seen in Equation 1.

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

- where TFij is the total number of occurrences of the word i in period j (considering all the documents), NNj is the total number of documents in period j, n is the number of periods and tw is a time weight. According to the proposal that new occurrences are the most relevant, the tw it has been defined as .05 by a group of experts (Yoon, 2012).

The "issue" dimension indicates the degree of diffusion of the subjects related to weak signals. This dimension is directly related to the frequency of occurrence of each word in each document, since this frequency is generally adopted to measure how general a term is in a collection of textual information (Salton and Buckley, 1988). To measure the Degree of Diffusion (DoD) of the word i in period j, Equation 2 is used.

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

- where DFij is the number of documents in which the word i appears in period j. In both signal and issue determination, future signals that have the possibility of being weak signals have, from a quantitative point of view, an absolute low occurrence frequency but a high trend fluctuation (a geometric mean of high DoD and DoV but a low occurrence frequency).

These equations provide the average of the increase ratios (DoD and DoV) of each word found in the complete set of documents analyzed, with respect to the frequencies obtained by years. With these data, two graphs are generated: the first is the "Keyword Issue Map", which represents a map of absolute occurrences of keywords in all the documents (related to DoD); and the second is the "Keyword Emergence Map", to represent the number of documents in which each keyword appears (related to DoV

Interpretation is an activity in which endosigns are formulated in the mind of an analyst, based on the exosigns of the issue. Based on this "interpretation" dimension of a sign, a new issue might emerge. As the first generated exosignals after a new discovery are usually the research results published in the form of an article in a scientific journal, this source has been used to measure the interpretation dimension of a sign.

All H-index values in the SCImago Journal portal for all the documents where a word i appears have been added to calculate the Degree of Transmission (DoT) of that word. This measurement is expressed by Equation 3.

$$DoT_i = \sum Hindex_{journal} \quad (3)$$

Then, all DoT values were normalized, and expressed graphically in both DoD and DoV graphs with a different size for each dot, where each one of the dots represents a different word. A big DoT indicates a higher transmission of that word.

3.7 Phase 5: Natural Language Processing

After the generation of these two graphs (with the addition of the interpretation dimension), a list of words is generated as weak signal candidates.

The next step is to determine the meaning of text selections such as word sequences, a process known as semantic analysis. To this end the context of the words identified as weak signal candidates needs to be analysed, to avoid the detection of many false positives. A multi-word expressions analysis is a deeper approach that will reveal more accurate information about the detected term of interest. Previous works such as (Tsenga et alii, 2007) and (Zhang et alii, 2008) show the importance of multi-word expressions as a text mining technique. For this study, a natural language processing tool, namely a multi-word analysis, was applied to the list of candidates for weak signals. The words immediately preceding and immediately following the word identified as a weak signal in all documents were considered. The combinations in which these additional words are in the list of standard stop words were discarded.

3.8 Phase 6: Interpretation and Evaluation

Finally, several outputs have been obtained from the execution of the model: (i) the Keyword Issue Map; (ii) the Keyword Emergence Map; (iii) the table of words detected as candidates for weak signals; and (iv) the multi-word study of the words detected in the previous step, providing more accurate information and discarding false positives. This information is given to experts, entrepreneurs, companies and other interested parties in the topic under study, to facilitate their decision making.

4 RESULTS OF THE SYSTEM APPLIED TO SOLAR PANELS AND ARTIFICIAL INTELLIGENCE

In this section, output results of the application of the implemented system applied to two different sectors will be described. The two chosen sectors are: (i) solar panels, and (ii) artificial intelligence.

4.1 Keyword Issue Map

As previously stated, words that are candidates as weak signals are topics that have an abnormal pattern but are rarely detected. From the perspective of quantitative analysis, these words are likely to have low absolute document frequency but a 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 important and well spread. The method computed the DoD of each word based on the document frequency of the word per news article.

DoT is also measured by the impact of the journals where the word appears. The keyword issue map is generated by using the average time-weighted increasing rate of document frequency in geometric mean, and the absolute average document frequency of each word. Figure 4 shows a graphic view of the Key Issue Map with the candidate words for weak signals in the first quadrant (for both sectors). The size of the dots corresponds with the DoT variable.

4.2 Keyword Emergence Map

The method computed DoV of each word based on the document frequency of the word per news article. DOT is also measured by the impact of the journals where the word appears. The keyword emergence map is generated using the average time-weighted increasing rate (geometric mean) and the absolute average term frequency of each word. Figure 5 shows a graphic view of the Key Emergence Map, with the candidate words for weak signals in the first quadrant (for both sectors). Table 1 shows an example with the word "animal", its stem, absolute frequency, increment ratio, DoV and its category.

Table 1: A keyword in the Key Emergence Map - Degree of Visibility, applied to the future of artificial intelligence.

Keyword	Stem	Frequency	IR	DOV	Category
animal	anim	6.55	0.1886	60.40	Biology

4.3 Detected Candidates for Weak Signals

If in both the Keyword Issue Map and Keyword Emergence Map graphs a word is detected, it means that this word fulfils the criteria to be related to a weak signal according to Hiltunen model. For the sector of artificial intelligence, there are 92 words appearing simultaneously in both maps. Some of them are: ability, adaptive, animal, biological, bottom, brain, building, care, clinical, code, cognitive, colony, competitive, conditional, cooperative, core, dataset, design, diagnosis, education, energy, experimental, forming, future, generation, growth, health, history, hybrid, identity, impact, inspired, internet, limits, making, nonlinear, particle, patient, point, potential, presence, principle, quantitative, rate, regression, review, role, scale, self, social, solution, transform, universal, well and yield.

For the sector of solar panels, 134 terms were related to weak signals. Some of them are: abundance, acceleration, acceptable, acids, adsorption, aerosol, Africa, agriculture, Ahmed, air conditioning, Algeria, algorithm, computation, concentrator, cycling, wavelengths, wheel and zinc.

4.4 Multi-word Analysis

A multi-word analysis was performed with these words to discard false positives and determine more accurate word expressions related to weak signals. Results in artificial intelligence show expressions related to weak signals such as "human brain", "artificial brain", "clinical clustering", "cooperative cooevolution", and "cooperative mobile", that can be investigated according to the categories detected previously in the study, for example "Biology" or "Neuroscience". These results are forwarded to the experts to help their decision making.

5 EVALUATION OF THE RESULTS

A previous work (Yoon, 2012) validated the results using the same set of documents which were the input of the system, so obviously, they confirmed their hypothesis. In other applications, a dataset is divided into training, test and validation sets. However, as weak signals have a low absolute frequency, a weak signal detected in the test set would rarely be also present in the validation set. This evaluation process was therefore discarded.

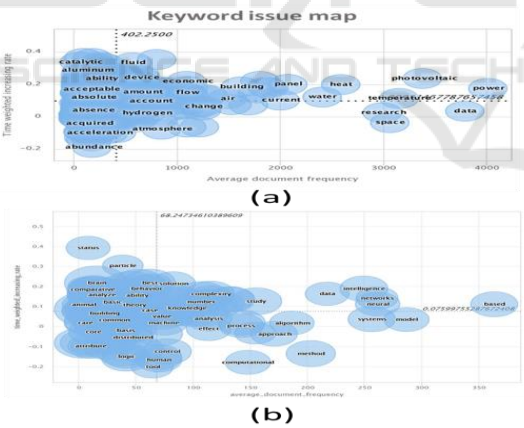

Figure 4: Keyword Issue Map (related to DoD) for (a) Solar Panels and (b) Artificial Intelligence.

Using different sources also provides a better mechanism to test the results, in comparison with independent results of documents from only a single source type. In this study, a ranking of the results taking issue and signal components, and another with the three components, show similar conclusions. Further studies should also determine the weight of different sources depending on the subject or sector to study. These weights can also change through time.

It has already been established that weak signals have an abnormal pattern. From the view of quantitative analysis, these words are likely to have low absolute document frequency but high range of fluctuation in the increase of document frequency as has been stated previously. Unlike previous works (Yoon, 2012) in which only web news is used, this study provides a new variable called Degree of Transmission based on the impact of the scientific papers as primary sources. The Hiltunen model explains that the interpretation component of the sign is also relevant in a weak signal. A multi-word expressions analysis shows more accurate and interesting information about the detected term of interest, than an analysis based only on single words.

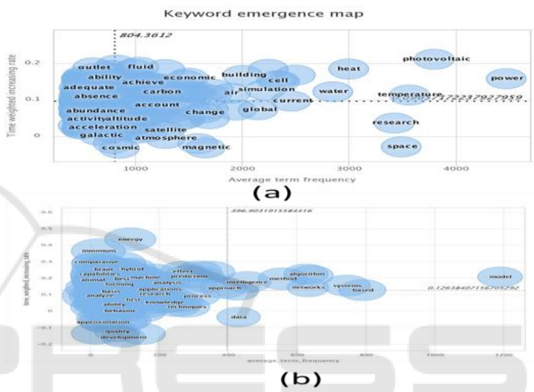

Figure 5: Keyword Emergence Map (related to DoV) for (a) Solar Panels and (b) Artificial Intelligence.

6 CONCLUSIONS

In this paper, a quantitative approach to design a system for detecting weak signals of the future is described, with an analysis for solar panels and artificial intelligence. In contrast to current work focusing on structured information sources, or using only a single type of data source, the detection of these signals is quantitatively based on heterogeneous and unstructured documents of various kinds (scientific journals, newspaper articles and social networks). In conclusion, this work suggests the possibility of quantifying the detection process of weak signals by text mining using multiple sources. The proposed method shows promising results to detect weak signals more efficiently than human experts when dealing with the massive textual information available due to the exponential increase in new documents related to any subject.

In future work, new topics will be added to evaluate the system. Moreover, to obtain more accurate results, other natural language processing tools should be added such as parts-of-speech tagging, bag of words recognition, regular expressions or sentiment analysis. In addition, hardware improvement will be implemented to obtain an even more efficient system.

ACKNOWLEDGEMENTS

This work was supported by the Department of Electronics of the Universitat Politècnica de València, IDEAS-UPV and the projects of Climate KIC and INNOCENS.

REFERENCES

Ansoff, H.I., 1975 Managing Strategic Surprise by Response to Weak Signals, in: *California Management Review*, vol. 18 (2), pp.21-33.

Bird, S., Loper E., Klein, E., 2009. Natural Language Processing with Python, in: *O'Reilly Media Inc*.

Brown, S., Eisenhardt, K. 1998. Competing on the Edge: Strategy as Structured Chaos, in: *Harvard Business School Press*, Boston, p.23.

Dator, J., 1996. Futures Studies as Applied Knowledge, in*: New Thinking for a New Millennium.* Richard Slaughter, ed., London: Routledge.

Dator, J., 2005. Universities without quality and quality without universities, in: *On the Horizon*, Vol.13, No. 4, pp.199-215.

Eckhoff, R., Markus, M., Lassnig, M., Schon, S., 2014. Detecting weak signals with technologies: overview of current technology enhanced approaches for the detection of weak signals., in: *Int. J. Trends Econ. Manag. Technol.* 3, pp.1-7.

Fischler, M.A, Firschein, O., 1987, Intelligence: The Eye, The Brain and The Computer, in: *Addison-Wesley*. p.221.

Godet, M., 1994. From Anticipation to Action, in: *A Handbook of Strategic Prospective, UNESCO Publishing*, p.59.

Guralnik, V., Srivastava, J., 1999. Event detection from time series data, in: *KDD '99 Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, San Diego, California, USA, pp.33-42. doi:10.1145/312129.312190.

Haegeman, K., Marinelli, E., Scapolo, F., Ricci, A., Sokolov, A., 2013. Quantitative and qualitative approaches in future-oriented technology analysis (FTA): from combination to integration, in: Technol. Forecast. Soc. Chang. 80, pp. 386 - 397.

Han, J., Kamber, M., 2001. Data Mining: Concepts and Techniques, in: *Morgan Kaufmann Publishers*, USA, p.236.

Hiltunen, E., 2008. The future sign and its three dimensions, in: *Futures*, 40(3), pp.247 - 260.

Inmon, W.H., 2005. Building the Data Warehouse, 4th ed., John Wiley, p.156.

Ishikiriyama, C.S., Miro, D., Simates Gomesa, C.F. 2015. Text Mining Business Intelligence: A small sample of what words can say, in: *Procedia Computer Science*, Volume 55, pp.261- 267.

Jung, K., 2010. A study of foresight method based on text mining and complexity network analysis, in: *KISTEP: Seoul*.

Khan, R.A., 2012. KDD for Business Intelligence, in: *Journal of Knowledge Management Practice*, Vol. 13, No. 2, p.134.

Kim, J., Han, M., Lee, Y., Park, Y., 2006. Futuristic data-driven sce- nario building: Incorporating text mining and fuzzy association rule mining into fuzzy cognitive map, in: *Expert Systems With Applications* 57, 311-323.

Koivisto, R., Kulmala, I., Gotcheva, N., 2016. Weak signals and damage scenarios Systematics to identify weak signals and their sources related to mass transport attacks, in: *VTT Technical Research Centre of Finland* Ltd, Tekniikantie 1, FI-33101 Tampere, Finland Technological Forecasting and Social Change 104, pp.180 - 190.

Kuusi, O., Hiltunen, E., 2007. The Signification Process of the Future Sign, Finland Futures Research Centre ebook 4/2007, p.7.

Mannermaa, M., 1999. Tulevaisuuden hallinta skenaariot strategiatyoskentelyssa. (Managing the future, Scenarios in strategy work), in: *Porvoo*, 1999, p.227.

Molitor, G.T., 2003. Molitor Forecasting Model: Key Dimensions for Plotting the Patterns of Change, in: *Journal of Future Studies*, August 8:1, pp.61-72.

Nikander, I.O., 2002. Early Warnings, A Phenomenon in Project Management, Dissertation for the degree of Doctor of Science in *Technology, Helsinki University of Technology*, 2002, p.196.

Peirce, C.S., 1868. Some Consequences of Four Incapacities, in: *Journal of Speculative Philosophy*, pp.140-157.

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

Thorleuchter, D., Van den Poel, D., 2015. Idea mining for web- based weak signal detection, in: *Futures* 66, pp.25-34.

Tsenga, Y., Lin, C., Linc, Y., 2007. Text mining techniques for patent analysis, in: Information Processing and Management, Volume 43, Issue 5, pp.1216-1247.

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

Yoon, J., 2012. Detecting weak signals for long-term business opportunities using text mining of Web news, in: *Expert Systems with Applications*, vol. 39, pp.12543-12550.

Zhang, W., Yoshida, T., Tang, X., 2008. Text classification based on multi-word with support vector machine, in: *Knowledge-Based Systems*, Volume 21, Issue 8, pp.879-886.