A Framework for Weak Signal Detection in Competitive Intelligence using Semantic Clustering Algorithms

Adil Bouktaib, Abdelhadi Fennan
LIST Department of Computer Science
Abdelmalek Essaadi University
Tangier, Morocco

***Abstract*—Companies nowadays are sharing a lot of data on the web in structured and unstructured format, the data holds many signals from which we can analyze and detect innovation using weak signal detection approaches. To gain a competitive advantage over competitors, the velocity and volume of data available online must be exploited and processed to extract and monitor any type of strategic challenge or surprise whether it is in form of opportunities or threats. To capture early signs of a change in the environment in a big data context where data is voluminous and unstructured, we present in this paper a framework for weak signal detection relying on the crawling of a variety of web sources and big data based implementation of text mining techniques for the automatic detection and monitoring of weak signals using an aggregation approach of semantic clustering algorithms. The novelty of this paper resides in the capability of the framework to extend to an unlimited amount of unstructured data, that needs novel approaches to analyze, and the aggregation of semantic clustering algorithms for better automation and higher accuracy of weak signal detection. A corpus of scientific articles and patents is collected in order to validate the framework and provide a use case for future interested researchers in identifying weak signals in a corpus of data of a specific technological domain.**

Keywords*—Competitive intelligence; apache spark; big data; weak signal detection; web mining; semantic clustering

I. Introduction

In the era of big data, information flows from different sources and in huge volumes. Companies and organizations are under many threats coming from different opponents and competitors. Strategic decisions must be made in order to survive the market changes and cultural, technological, or political shifts that may occur in their environment. Economists rely on the most popular models for strategies to conduct a thorough competitive intelligence activity [1][2] for example : SWOT analysis's main purpose is to analyze threats and opportunities and develop plans to react strategically to those events, this model can be supported by using weak signal detection and early warning signs techniques [3].While PETS model analyzes the data concerning the environment of the company by monitoring political, economic, technological and social factors in order to prepare strategic responses to any change so it can maintain a dominant position in the market. Many organizations invest heavily in developing systems to automate the process of competitive intelligence [4] [5] and implement their adopted strategies. One of the main goals and features of those systems is the detection of weak signals in the environment surrounding an organization. Environmental scanning is gaining the attention of many stakeholders due to the benefits and advantages [6] it brings to the well-being and the contribution to the sustainability of their companies. The aim of most environmental analysts is to detect pieces of valuable information that will give them the strategic advantage of anticipation and early response planning, this can be done through weak signal detection. A weak signal is defined as a temporal change that occurs in a domain or a topic or in the environment in general [7], and it may have an impact on the future and become a trend. Therefore, the early detection and identification of this strategic information is crucial to the evolution of an organization. Many definitions are given to this concept, and different techniques and approaches are applied to detect this kind of information automatically, which is the subject of the next sections.

Companies must be able to understand and explore their environment to extract implicit knowledge that cannot be identified by experts. But it should also be able to predict the future evolution of a specific domain. The emergence of web data and the availability of information online pushes the companies nowadays to exploit these data to extract meaningful strategic information that allows them to make optimal and strategic decisions based on a scientific accurate analysis of the data, and an intelligent approach of web mining [8] to extract high-quality data.

Competitive intelligence systems are software that groups together a set of tools and technologies that companies have to implement in order to keep track of their evolving environment [9]. Many of these solutions neglect the anticipative information model that helps predict and monitor trends that unfold threats and opportunities that must be harnessed and used to gain a competitive advantage.

Weak signals are pieces of information that will help companies to identify threats and opportunities in their environment, which in turn will allow the implementation of an anticipative strategy rather than a reactive one that responds to the events as they happen.

With the rapid stream of data available online and the vast number of documents available on the web every second,

companies must use the latest big data technologies and advanced algorithms in order to process and analyze this data efficiently to identify weak signals [10]. In this paper we propose a framework for weak signal detection in collected data from the web, using big data technologies and aggregation of semantic clustering algorithms based on Apache Spark to detect weak signals and emerging trends and monitor opportunities and threats.

This paper is structured as follows: in section 1, we present the definition of the main concepts of this work: competitive intelligence, SWOT analysis strategy, weak signals detection, competitive intelligence systems, big data analytics, semantic clustering algorithms. Section 2 will present some of the related works that try to handle and propose novel tools and methods of weak signal detection and we will highlight some of their limitations. Section 3 presents the proposed framework and our approach to detect weak signals. Section 4 presents the results of a case study in collected articles about "big data", and we show the results of our approach, then we finish by a discussion and conclusion.

II. Problematic

In order to monitor competitors and identify early warning signs that help decision makers identify companies' key intelligence needs [11], we need a framework for weak signal detection that will allow us to listen to and anticipate the changes in the market [12] by providing an unsupervised manner of analyzing data and capturing potential weak signals that evolve through time.

We define the problem and the importance of our contribution as follows: The main problem is how to process and analyze large amount of unstructured big data automatically from various sources to detect weak signals and unveil some of the strategic information hidden in a large corpus of textual documents.

We use semantic clustering algorithms with an aggregation approach to automate the detection of weak signals that share some characteristics that we defined earlier in the framework and we propose them to the final user domain expert who will then judge their usefulness in a strategic decision or action.

Most solutions do not process a variety of sources and big data, so we will try to propose a framework that is capable of analyzing data coming from multiple sources, and architecture to support the evolution of volume and velocity of data while relying on Apache Spark capabilities and semantic clustering algorithms like LDA (Latent Dirichlet Allocation), LSA (Latent Semantic Analysis) and K-Means[13] to give accurate results and high semantically related clusters of terms that may represent a weak signal.

III. Main Concepts

A. Competitive Intelligence

Competitive intelligence is defined as a process, activity, service[14] that starts from the definition of a strategic need problem, passing by the collection of multiple data from different sources, and through the analysis of this data, analysts process the data using their set of tools and techniques to extract strategic information from the data and interpret it to transform it into a usable and a valuable knowledge to be disseminated to the stakeholders, every organization has a different model and strategy to conduct competitive intelligence, which varies depending on the size, the environment or the need of an organization, in order to enhance the decision making process.

The goal of conducting competitive intelligence is to define the position of an organization in the market and to help it be aware of the changes and competitive forces around its environment [15], by providing an organizational tool capable of generating valuable knowledge from raw data to guarantee better business performance by taking strategic actions at the right time [16].

B. SWOT Analyisis and PEST Model

Many models exist to implement competitive intelligence monitoring strategies. Economists proposed models to establish an environment scanning tools to prepare for any change in the market and give an objective perspective of the position of an organization. SWOT analysis [17] focuses on analyzing the strengths and weaknesses of a company through processing internal data, and opportunities and threats coming from external data. When talking about weak signals, we are more interested in analyzing the opportunities and threats coming from the market. The PEST model [18] stands for political, economic, social, and technological factors of an environment that could influence the existence of an organization and its evolution in the market. That external information can be collected and analyzed easily from external web data and exploited in technological intelligence to be able to detect innovation [19], which is present in both structured and unstructured form. The aim of this paper is to analyze big unstructured data using big data analytics technologies and efficient algorithms while respecting and following the main ideas and concepts of those two models, as in Fig. 1.

C. Weak Signals

According to Igor Ansoff [20], weak signals are defined as small changes and imprecise early indications that occur over a period of time on a specific topic that may have an ongoing impact on the future. Weak signals are temporal changes that hold important and strategic information that companies and organizations must detect and collect to stay ahead in the market [21]. This helps them implement an anticipative approach of handling the opportunities and threats present in the market in the form of unstructured data harvested from the web.

The identification of weak signals relies on some characteristics and key points. According to Ansoff weak signals are weakly mentioned in their first appearance, they are less frequent than the main concepts in the context where they exist, but they are new and novel and hold a sign of innovation or a surprise in the market. The interpretation of weak signals requires domain experts in order to contextualize the findings and transform data into knowledge, and classify them as opportunities or threats and disseminate them to stakeholders to make an appropriate strategic decision.

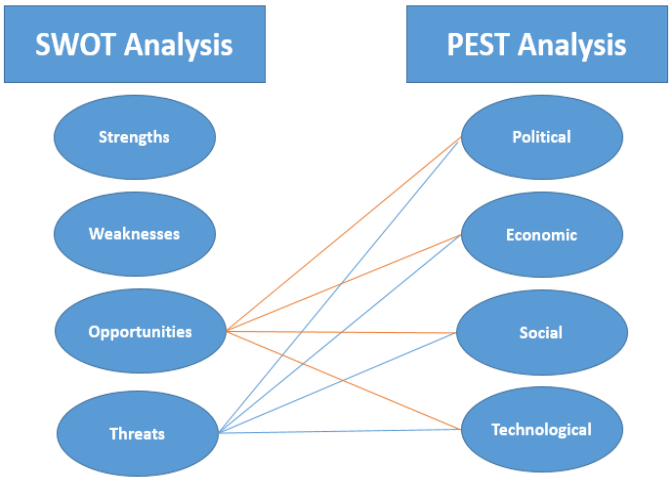

Fig. 1. SWOT Analysis vs PETS Model.

D. Apache Spark

Due to the volume of data available online, data must be collected from different sources in different formats. A homogenization step is mandatory to unify the structure of the data to be collected. Once the data is collected, we end up with huge volumes of data that cannot be processed by a normal computing approach, thus the need for big data analytics technologies that support huge volumes and fast streams of data. Few weak signal detection researchers have proposed a technological framework that addresses the issue of big data. Therefore, we propose in this paper a big data framework for weak signal detection with the implementation of semantic clustering algorithms in Apache Spark.

Apache spark [22] is one of the main big data analytics technologies, and the most well-known platforms for massive distributed computing, that are popular nowadays. This framework is gaining a lot of attention in the big data community and its use in a variety of applications [23] proved to give efficient results when dealing with large datasets. Hence we chose this framework in our attempt to develop a competitive intelligence system [24] to analyze and extract strategic information from the increasing amounts of data available in the environment of companies and organizations.

Apache spark has been used in a lot of applications [25] and it has been used to implement a variety of big data platforms and solutions. Apache Spark is a part of the Hadoop ecosystem introduced in 2009. While Hadoop processing is based on the MapReduce computing paradigm, Spark relies on the DAG (Directed Acyclic Graph) paradigm, which imposes sequential processing of RDDs, a distributed unit of data nodes in the cluster, that optimizes the consumption of resources by avoiding costly data copies used in iterative algorithms that we are going to be using in the weak signal detection framework.

IV. Related Work

Many researchers tried to apply variable methods to detect efficiently weak signals in large volumes of documents [26]. Those methods vary from supervised to unsupervised machine learning methods, automatic and semi-automatic methods, or manual methods relying on experts input, quantitative and qualitative methods, and many data sources were used to prove the approaches and detect weak signals.

One of the early approaches and attempts to discover weak signals in data, Yoon [27] used a keyword-based text mining method to identify opportunities in web news data. He used a quantitative method in which he performed a time-weighted analysis by calculating the occurrence and frequency of keywords during a period of time. But this attempt was limited to only one source of data, and it lacks an automatic crawling of data from multiple sources, and fails when it comes to dealing with large datasets. The result may not be easily interpreted when visualizing a large space of keywords.

El Hadadai.Anass et al [28] proposed a sequence data mining based method for extracting emerging trends and highlighting the evolution of domains through crossing terms with dates and other fields. With the application of correspondence analysis and multiple correspondence analysis and a visualization tool, this approach was able to extract clusters of weak signals from sorting and extracting clusters from the obtained matrix. The method was evaluated using a dataset from scientific articles and patents collected from scientific databases in order to identify technological weak signals, but this method lacks the possibility to be extended to support large datasets and its need for an expert to manipulate the tool to perform the analysis.

D. Thorleuchter et al [29], proposed a methodology based on idea mining and Latent Semantic Analysis to identify weak signals, by constructing a matrix based on the vectors and patterns discovered from the idea mining approach, by applying a dimensionality reduction on the matrix using SVD decomposition, which produces a set of semantically related clusters that may be a weak signal. The method is limited, as stated by the authors. They observed that the method lacked the possibility to discover implicitly cited weak signals and proposed an enhancement using Latent Dirichlet Allocation to get more accurate results.

Antonio.L.et.al [30] proposed a method to conduct an anticipative intelligence by analyzing text and identifying weak signals, using clustering k-medoids and a Jaccard function as a similarity function between obtained clusters in order to analyze similar clusters of weak signals, the method claims to be automatic but the dataset is collected from experts at the beginning of the process.

Julien Maitre et.al [31], the work that is closely related to what we are proposing is inspired by this paper, which presents a novel approach for weak signal detection in weakly structured data or unstructured data, by combining Latent Dirichlet Allocation and Word2Vec algorithm to perform clustering on a corpus of documents collected from the web, the article proposes also a method to identify the number of clusters k to be extracted from a corpus using LDA, which in most cases is hard to define and is crucial to the quality and robustness of the obtained results especially when it comes to weak signals, where the use of a small k may eliminate the identification of important weak signals.

In our approach, we try to group the three algorithms in order to reduce the mistakes and weakness of those approaches

by using a clustering aggregation [32] approach supported by the computational power of Apache Spark and the flexible nature of RDDs and their reusability in iterative algorithms in order to perform multiple tasks, and with using the ML pipeline feature of Apache Spark to facilitate and automate the process of weak signal identification with a minimum interaction of experts.

V. Proposed Framework

In light of the findings of the literature review conducted by C. Muhlroth et.al [26] and other reviewed approaches [33] [34] [35] [36], we found a need to propose a big data analytics framework for automatic weak signal detection. Thus we propose in this paper a framework that uses Apache Spark to implement the data analysis from data collection to weak signal identification using semantic clustering algorithms. The feature of Apache Spark that allows us to achieve this is the ML pipeline that aims at automating steps to be applied on a dataset, in order to extract implicit hidden information that may present key strategic indications to be processed and analyzed.

Fig. 2 presents the architecture of the proposed framework. It outlines the steps to be followed in the pipeline implemented using Apache Spark, starting from data collection to the identification of weak signals contained in the corpus of collected documents. In the following section, we provide a brief explanation of Apache Spark ML pipeline, and we explain the steps of the pipeline in detail.

A. Apache Spark DAG and ML Pipeline

Apache Spark provides an API to manipulate RDDs, resilient distributed datasets, which is a good structure for dealing with big unstructured data. The power of this data structure remains in the possibility to expand to huge volumes of data, thus the adoption of this technique in our work. RDDs will hold the corpus data to perform analysis using ML pipeline API that represents a set of processes to perform on a dataset to get the desired results. This makes it easier to aggregate multiple algorithms into a single pipeline. We will be using this technique in our work to implement an efficient big data analytics framework for weak signal detection, by combining the semantic clustering algorithms presented in Fig. 2.

The technology that allows Apache Spark to execute such processing is DAGs which is a new enhanced strategy to perform map-reduce tasks, as shown in Fig. 3, by organizing the planning of execution in stages and steps that form a directed acyclic graph of transformations to apply on the dataset.

All the algorithms used in this framework will be implemented using the Apache Spark MLlib library that contains a variety of tools and machine learning algorithms and clustering to be applied on the data. We combine LDA and implement LSI and K-means by using ML Pipeline to perform semantic clustering on the corpus of collected data. At the end, we communicate the findings to the stakeholders and experts to identify the clusters that hold potential weak signals.

B. Data Collection

The framework starts with data collection. We collect data from multiple scientific articles databases and patents and store them in the Hadoop file system. When we start the execution of the ML pipeline we load the data from Hadoop onto the Apache spark cluster in order to execute the outlined pipeline process depicted in Fig. 2. Scientific databases from IEEExplorer, ACM Digital, and patents from USPTO, contains many articles and documents having a lot of fields like text, date, abstract, publication date, etc. We are interested in the text and publication date of the document in order to conduct a temporal analysis of weak signals and the evolution of the topic in time to perform technological surveillance on a specific field of interest.

Many scrappers and crawlers are developed to collect data from those websites using web mining methods and Scrapy python framework [37], which gives the possibility to create scraping agents to crawl as many web pages as possible with the elimination of repeated documents. Our approach helps decision-makers and analysts to collect data automatically and conduct environmental scanning with no need for manual intervention, which could be a hard task for companies in this era of big data.

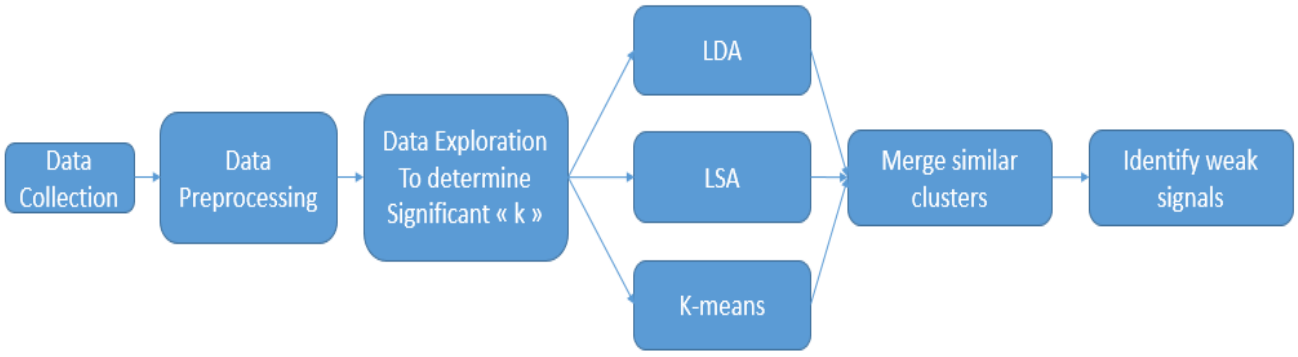

Fig. 2. Proposed Framework Architecture.

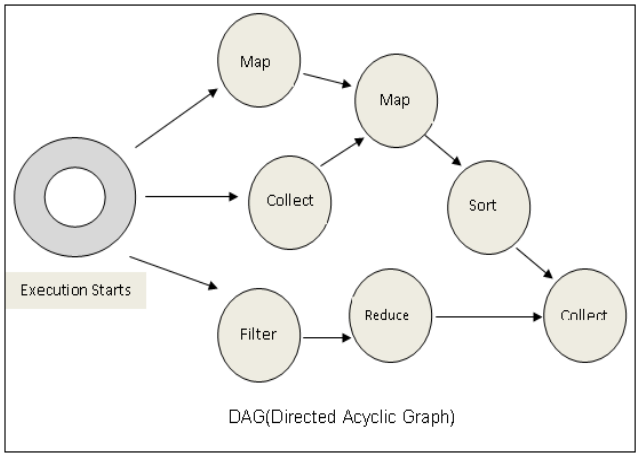

Fig. 3. DAG Excution Method in Spark.

C. Data Preprocessing

Preprocessing is an important step to clean data and format it to our needs. After choosing the text field we will use in analysis and the date field that will help us to filter emerging trends, we clean the text from ambiguous characters, then by removing StopWords, stemming and lemmatization, which will help us to get more accurate results and easily interpretable information from raw data. We create n-grams from the corpus to add them to the vocabulary of the corpus. This step is important to enable the clustering algorithms to identify multi-terms that may hold an important part of a weak signal, especially in the scientific field. Due to the nature of weak signals, which is low frequency and occurrence of words, we eliminate terms where the count is above a threshold, for example 200 occurrences, as we are not interested in highly frequently mentioned words that, in most cases, represent strong signals or trends, which are not the purpose of our analysis.

D. Data Exploration

The number of topics to be extracted cannot be determined previously as the algorithms used are unsupervised algorithms and the analyst does not have an idea about the number of clusters to be obtained. Therefore, we choose a rule of thumb and we define the number of clusters to be extracted as in eq.1 after the extraction of the vocabulary from the corpus:

$$k = \sqrt{n/2} \quad (1)$$

where n is the number of words in the vocabulary of the corpus.

The determination of an approximate k is an important step in the process of this pipeline. We can specify k based on many techniques of data exploration or using many methods from the literature [38], which is outside the scope of our research, or we can try different values of k and analyze the different clusters obtained. A small number of k though must be avoided in order to avoid the elimination of important potential weak signals that are not heavily cited.

E. LDA

Latent Dirichlet Allocation [39] is a generative probabilistic model for text classification and a topic modeling algorithm that aims at representing the documents as a set of topics, with the objective of assigning each term to a semantically related topic. When applying LDA in Fig. 4 to a corpus of documents, the algorithm tries to cluster the topics and their related terms according to their semantic relationships. It identifies k topics, k is a number specified by the analyst, many methods exist to choose the best k that gives accurate clusters.

LDA algorithm steps are defined as follows, for each document w in a corpus D:

1. Choose N ~ Poisson(ξ).
2. Choose θ ~ Dir(α).
3. For each of the N words wn:
   (a)Choose a topic zn ~ Multinomial(θ).
   (b)Choose a word wn from p(wn |zn,β), a multinomial probability conditioned on the topic

Those steps are the standard for the LDA model, in order to cluster a distribution of semantically related words to a set of specific topics, in our case those topics may represent innovations, opportunities or threats.

So we will use this algorithm to detect underlying topics in a corpus of documents. Those topics may include weak signals that are not easily identified and are not in the scope of the knowledge of experts, especially in the case of new innovations in a domain. After removing the most frequent terms from the documents, we aim at identifying sets of words that are less frequent and semantically related and belong to the same topic. That's why a maximal number of k is essential to the extraction of latent topics that represent a small proportion of the document, which is the nature of weak signals defined by Ansoff.

F. LSA

Latent semantic analysis [40] is a text mining technique that aims to create a semantic space to identify relationships between words in a corpus of documents. Those relationships are semantically detected using a linear algebra technique called SVD decomposition. Its goal is to decompose a document-term matrix created from the corpus into a lower-dimensional space in order to detect close words and extract coherent topics and similarities between documents.

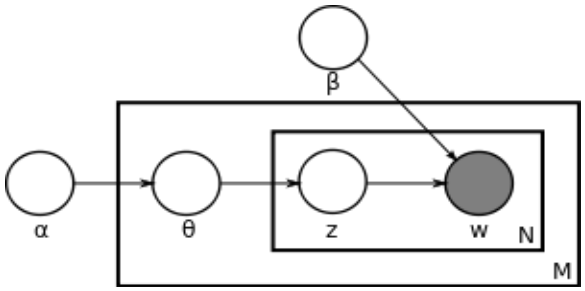

Fig. 4. Plate Notation Visualizing the LDA Model Parameters, Plate M Represent The Total Number of Documents, N Represent the Numbers of Terms in a Document, α the Per-Document Topic Distributions, β the Per-Topic Word Distribution, θ the Topic Distribution For Document m, z the Topic for the n-th Word, w is a Specific Word.

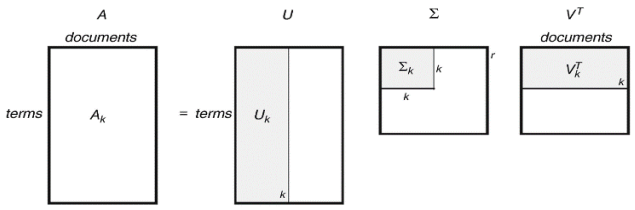

Fig. 5. LSA Algorithm Matrix Decomposition Process.

By applying this technique in weak signal detection we want to detect weak clusters that are appearing in the corpus, those highly coherent and newly emerging clusters may hold important strategic information, they may represent an opportunity for investment and collaboration, or a threat that a company has to plan a strategic response to face it and overcome its consequences.

After the creation of the matrix from the collected corpus by crossing the terms with their corresponding documents, we create an m*n matrix where m is the number of terms and n is the number of documents, then we apply SVD which will decompose the matrix into 3 new matrices as depicted in Fig. 5.

$$M = U\sum V^* \quad (2)$$

Where U is an m*k matrix that holds the word assignment to topics, Σ an k*k matrix which contains singular values that represent the importance of the topic, V* is an k*n matrix that contains the topic distribution across documents. In our case, we are interested in the first two matrices. By crossing the pairs of vectors of the two matrices, we obtain the clusters of topics and their corresponding terms. The clusters obtained in this step will be merged with the previous results to enhance the semantic understanding of the corpus.

G. K-MEANS

K-means is one the most important algorithms for clustering data, the power of this approach resides in its ability to perform unsupervised learning and clustering of data with no prior knowledge, hence the choice of this algorithm in our process to enhance the results of our approach and the quality of the obtained clusters.

As in the previous algorithms, we perform data preprocessing depicted in the previous section before applying k-means. To apply k-means we follow those steps in order to find semantically related terms, relying on the Word2Vec [41] model, and group them in cluster as depicted in Fig. 6:

- Cleaning and Preprocessing of text.
- Determination of number k.
- Feature extraction using Word2vec to represent each word semantically as a vector.
- Applying k-means.
- Getting clusters.

In the next step, we will merge the most similar clusters to unify and expand the clusters that are candidates to be weak signal clusters, containing information about potential opportunities or threats that must be noticed.

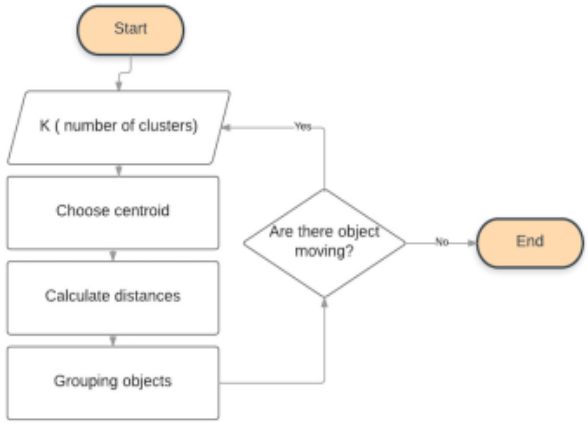

Fig. 6. K-Means Algorithm Steps.

H. Cluster Aggregation

Cluster aggregation [32] is a method that aims to apply different clustering algorithms on the dataset and find a consensus about the optimum cluster groups in order to eliminate duplicates, and eliminate the noise of each algorithm if it was applied individually, in order to improve the quality and robustness of the clustering.

After extracting the clusters from the previous steps, denoted C1, C2 and C3 from applying LDA, LSA and K-means respectively, we move to the merge step which consists of performing a similarity calculation between all pairs to identify similar clusters and merge them in order to eliminate redundancy and enhance the quality of the weak signals detection process by minimizing the disagreements between clusters according to Equation 4.

$$D(C) = \sum_{i=1}^{m} d_V(C_i, C) \quad (3)$$

where v is a set of words or multi-terms and m is the number of all clusters from the applied algorithms.

The implementation of Approximate Similarity Join of Apache Spark MLLib is used, which is based on the Jaccard similarity function eq (4). We calculate it for each pair of all clusters from all algorithms, and if it passes a threshold, we merge the clusters into one, in order to get the cluster that minimizes the number of disagreements.

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \quad (4)$$

The resulted clusters from Algorithm 1 are shared with experts and stakeholders to identify potential weak signals from the corpus.

I. Weak Signal Identification

Extracted clusters will pass by the last step, which aims at calculating a score that represents the weighted term evolution inspired by Yoon et.al [27], the evolution rate "er*" of each term during a period t of the cluster "Ci" is calculated and the sum "eri" of all terms represent the score of a cluster, based on that score we can identify the clusters that may hold weak signals represented by semantically related terms from the corpus.

We order the clusters by their score, and based on that score and the interpretation of an expert in a domain, we can spot the clusters that are holding information about the weak signal, which, by interpretation, may be a threat, an opportunity of investment, or an innovation that needs further investigation or collaboration.

In the next section we present the results of applying this flow to the collected corpus, and we discuss the obtained results, advantages and limitations of our framework and we conclude with ideas for future researchers.

```
Algortihm 1: MergeClusters
Input: clusters [C1, C2, C3]
Output: megred clusters [Cm]
BEGIN
    For each pair of clusters:
    Calculate similarity
    If similarity > threshold
        DO merge_clusters ()
RETURN merged_clusters
```

VI. Result and Disccusion

In order to evaluate the proposed method, we will conduct an analysis on a dataset of scientific articles about "big data" topic, we collected a corpus of 5800 documents and scientific articles about « Big Data » containing multiple fields, from the fields we are interested in are abstract field and publication date. We will perform text mining on the text field and perform growth analysis using the publication date.

The purpose of our analysis is to perform the clustering aggregation of three algorithms, K-means, LDA, and LSA in order to combine the results of each algorithm and select from the obtained clusters the ones that are potential weak signals and may hold information about opportunities or threats.

A. Data Collection

We collect data from IEEExplorer, ACM Digital Library, SpringerLink and Sciendirect to show a case study and illustrate the processes of our approach. We use the search query "big data" and choose a publication date range from 2000 to 2020, then we collect the documents and articles published in this range of time. We are interested in the abstract, title and publication year of a document as in Fig. 7, the scraper agents extract those fields using the CSS styling of each database website in order to ease the step of homogenization of those fields in the next step.

We create a data frame from the documents containing the three fields we are interested in, and process them in the remaining steps of the framework. Fig. 8 shows an extract of the collected data.

Fig. 7. IEEEXplorer Document Example.

Document Title	Abstract	Publication Year
A survey of data ...	Computer clusters...	2020
A Novel Data-Driv...	Data-driven appro...	2018
A novel clusterin...	Big data analytic...	2019
A Framework for B...	The emergence of ...	2019
Mining conditiona...	Current Condition...	2020
Social Set Analys...	Current analytica...	2016
Big data oriented...	Due to the tremen...	2018
Big Data, Big Kno...	The idea that the...	2015
Protection of Big...	In recent years, ...	2016
Big Data Analytic...	Mobile cellular n...	2016
Big Data Analytic...	Big data analytic...	2019
Big Data-Based Im...	Big data-based ac...	2019
An Integrated Met...	The expand trend ...	2019
A Methodology of ...	The traditional b...	2018
Analysis and Visu...	With the developm...	2019
Evaluating the Qu...	The use of freely...	2015
A Novel Online an...	A sizable amount ...	2017
Data Lake Lambda ...	The advances in s...	2018
Big data analytic...	In recent years, ...	2019
A Big Data Mining...	In recent years, ...	2019

only showing top 20 rows

Fig. 8. Dataframe Collected from the Scrapping Agents of the Framework.

B. LDA Obtained Clusters

After the preprocessing and cleaning step of the articles obtained about big data, we perform the first clustering algorithm, LDA, to obtain k clusters from the corpus. The topics obtained are semantically related and clustered in one group.

A sample of clusters obtained from the corpus is presented (Table I):

TABLE I. List of Obtained Clusters from LDA

Cluster	terms	
Topic1	new concept secure communication collection data imbalanced data method consists probability distribution time big	experiments carried nir fmt proposed algorithms location data error rmse minimize total intermediate pointers
Topic2	low power performance overhead set data model predict clustering method method paper features paper	non uniform information big learning architectures cold start tasks cloud signal quality address challenge
Topic3	traffic data information extraction conducted evaluate key challenges network operators computational efficiency selection strategy	characteristics data secure communication model based factors influence change detection processing time human interaction
Topic4	important information processing techniques video analysis word embeddings enhance performance deep hashing physical systems	linear programming existing work data intensive iot networks control mechanism stream processing cover problem

C. LSA Obtained Clusters

The application of latent semantic analysis is done. After applying LSA we obtain a different set of k clusters using the matrix decomposition of singular values (SVD). For each cluster, we select a set of terms that represent this concept and that are closely related to it using the singular values in the sigma matrix.

By applying the LSA on our corpus of data we obtain the following clusters (Table II):

TABLE II. LIST OF OBTAINED CLUSTERS FROM LSA

Cluster	terms	
Topic1	defect detection data fusion sar image detection method power supply tree boosting digital twin	high performance smart manufacturing data digital key management parking lot fabric defects address problem
Topic2	trajectory data lane changing social networks data driven data processing deep neural reinforcement learning	changing model spatial temporal uav bss onset date modis derived brain health health quality
Topic3	proposed model extensive experiments multiplicative linguistic uncertain multiplicative location privacy decision making city brain	cloud computing compared state massive datasets data processing experiments conducted communication consensus group decision
Topic4	point cloud ant colony time delay electric power neutrosophic cubic point clouds dense point	security threats power data uwan security power systems algorithm based cloud generation healthcare insurance

D. K-Means Obtained Clusters

The application of k-means results in a set of k clusters after the calculation of word2vec of the text to create a feature of semantically related words. This was used as the measure of similarity between words or terms to perform semantic clustering. The following clusters were obtained from applying this algorithm:

In the next step, we will try to merge similar clusters into one cluster and build a cluster group that collects the power of all the algorithms and solves the problems and weaknesses of the other approaches (Table III).

TABLE III. LIST OF OBTAINED CLUSTERS FROM LSA

Cluster	terms	
Topic1	big personal personal data hidden transition process model industrial internet open data redundant tight	data big results indicate data processing attack path subgraph matching smart cue value data
Topic2	health big data attracted data frameworks process big computing data processing architecture analysis big dimensional big	industrial big data present processing analysis hot research data techniques lte network statistical analysis data problem
Topic3	network models network model networks cnns based convolutional recurrent neural learning based	deep convolutional compared state vector machine networks cnn outperforms state network based'
Topic4	data analysis bda applications data collected paper present smart cities applications cloud incomplete information	rare events wireless networks public key things iot based data driving range entropy loss

E. Aggregation Algorithm Obtained Clusters

By applying the approximate join similarity, we get the pairs of similar clusters, by merging similar clusters we get p clusters $p < k*3$, which gives an idea about overall clustering and solve the mistakes that could have been made by using one individual algorithm, the obtained clusters represent all the small topics and semantically related terms that may hold an opportunity or a threat (Table IV).

TABLE IV. IDENTIFIED SIMILAR CLUSTERS

Cluster id	Cluster id	Similarity distance
33	30	0.307
36	33	0.307
23	36	0.428
25	6	0.428
23	1	0.428
6	25	0.444
36	23	0.444
23	24	0.461
24	23	0.461
1	23	0.473

We merge similar clusters into one in order to eliminate redundant clusters and improve the quality of visualization.

In order to visualize the results of the approach, we create a graph from the adjacency matrix term-topic and plot the graph using Gephi to see the clusters and the relationships between them. The graph obtained contains 670 nodes and 12 006 edges. We show an extract of the graph in Fig. 10, and an identified weak signal in Fig. 9 containing semantically related words about the application of big data in health.

F. Interpretation and Discussion

From the interpretation of the results, we can spot the weak signals and hidden information that are not visible to the experts, and by combining their expertise with results obtained, we can identify clusters that are potential strategic information holders and we should cross the data back to the original document for further analysis and understanding of the context of appearance and the identification of the importance of the discovered piece of information.

In our approach we filtered weakly cited words in a specific time, year of publication, from the corpus and applied three semantic clustering algorithms in hope of finding the most accurate clusters by using an aggregation method. Those obtained clusters may contain pieces of information that is crucial to the implementation of an anticipative strategy of an organization. A weak signal is characterized by the evolution of its presence or its number of occurrences through time, which makes it a strong signal in the future, though not all weak signals are destined to be strong.

In Fig. 9 we present the graph representation of filtered words from the corpus, those words are related by their co-existence in the same document and their appurtenance to the same cluster. In Fig. 10, we singled out a cluster so we can study the semantics of this potential weak signal with the help of a domain expert.

We see in Fig. 10 that the semantic cluster of topic 32 is weakly cited and highly rated in the last period of research, which means that this low visibility cluster may be a trend in the future, though we can comment on the choice of number k, which must be chosen wisely and we must experiment with different values of k, or we have to use a different algorithm to determine the optimal value of k that will give promising and accurate results.

Extracted Potential Weak signals must be harnessed to identify threats and opportunities in the market. Our method extracts the most promising clusters of weak signal topics. Using our approach and with expert intervention, we can spot the key information that will generate value for organizations. Though the advantage of this method is not to predict which weak signal will become strong, but to enhance the quality of extracted clusters from the corpus, so we can keep and analyze only the semantic clusters holding potential weak signals through the aggregation of three algorithms: LDA, LSA, and k-means, this approach will not eventually predict which one will become strong in the future. In order to predict whether a weak signal will become strong, we require labeled data, and with the application of supervised machine learning [42], we can extract the features of weak signals that are candidates to be strong and trend in the future.

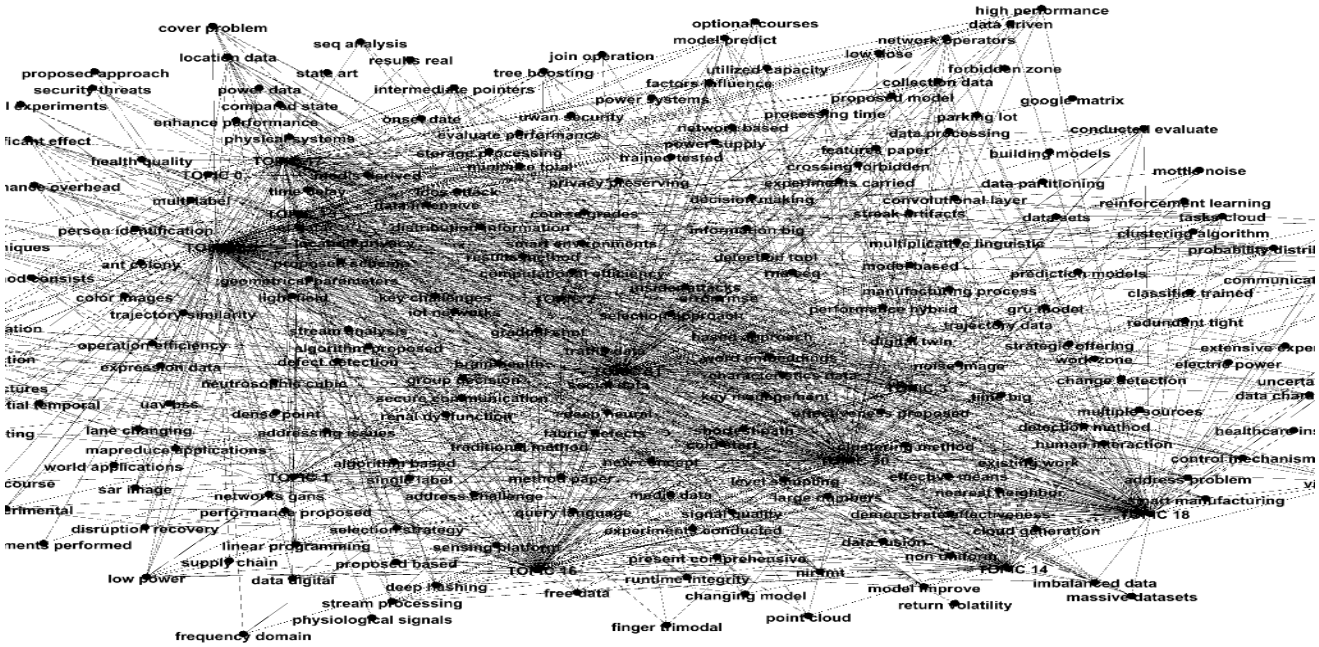

Fig. 9. Network of Collected Data from the Scrapping Agents of the Framework.

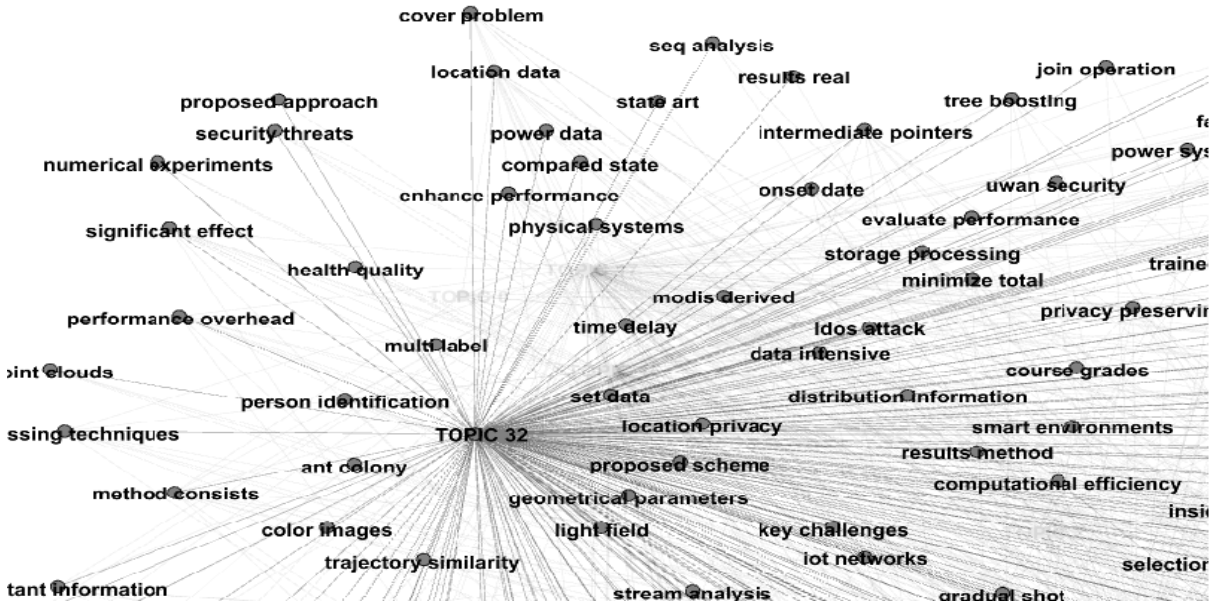

Fig. 10. Illustration of an Extracted Weak Signal Topic to be Analyzed (Topic 32).

VII. Conclusion

In this paper, we proposed a novel framework that uses semantic clustering aggregation models, made possible with the use of the computation power of Apache Spark, through the use of ML pipeline which gives the possibility of automating the process of weak signal detection in large volumes of data. The aim of aggregating clustering methods is to combine the features extracted from each method and hide its weaknesses. The use of such a tool in business will allow businesses and stakeholders to remain active and alert in the market. Semantic clustering methods have proven to be very efficient when it comes to topic modeling and extraction of variable topics semantically related from a large corpus of documents. Hence, we judge this approach to be very pragmatic in nature and it does contribute to the domain of weak signal detection.

Our framework will help stakeholders identify and prepare scenarios of intelligence needs. For each potential weak signal, there must be a strategic response ready to tackle it, which will help stakeholders implement an anticipative approach to conduct strategic competitive intelligence in a big data context, where manual extraction and analyzation of documents is impossible in this era where new data is available every millisecond.

Despite our framework contribution in the field we still think that there is more work to be done for future researchers in weak signal detection literature, for example the nature of data to be analyzed in weak signal detection research is unstructured, thus the need for more advanced clustering methods to perform unsupervised machine learning to label data as weak signals from the past data, and apply Text mining Deep Learning models [43] in order to be able to extract and identify weak signals in future data once available online, which will give a competitive advantage for organizations. In our future work we will apply Graph embedding technique [44] [45] as a technology that will allow us to reduce the dimensionality of the corpus and facilitate the semantic representation of weak signals, through the study of dynamic graph embedding to monitor the evolution of a domain terminology through time, in hope of detecting innovation, opportunity or a threat as early as possible.

In conclusion, we must mention the limitation of methods and approaches to validate the extracted weak signals in most of the literature [46]. As a future research project, we can propose a new direction of research in this field through adopting novel semantic clustering algorithms that rely on deep learning like Word2Vec and Glove word embedding [47] for a more precise semantic analysis of the corpus, and proposing novel approaches that relies on labeled data.

References

[1] Fleisher, Craig S., and Babette E. Bensoussan. Strategic and competitive analysis: methods and techniques for analyzing business competition. Upper Saddle River, NJ: Prentice Hall, 2003.

[2] Prescott, J. F., & Miller, S. H. (Eds.). (2002). Proven strategies in competitive intelligence: lessons from the trenches. John Wiley & Sons.

[3] Popescu, F., & Scarlat, C. (2015). Limits Of Swot Analysis And Their Impact On Decisions In Early Warning Systems. SEA: Practical Application of Science, 3(1).

[4] Sauter, Vicki L. "Competitive intelligence systems." Handbook on Decision Support Systems 2. Springer, Berlin, Heidelberg, 2008. 195-210. DOI: 10.1007/978-3-540-48716-6_10.

[5] Echternacht, Tiago Henrique de Souza, et al. " Competitive Intelligence: A Study Of The Involvement By Senior Corporate Management At Redepetro Companies." (2020), UFRN Electronic Journals Portal. http://rebacc.crcrj.org.br/handle/123456789/5179.

[6] Babatunde, Bayode O., and Adebola O. Adebisi. "Strategic Environmental Scanning and Organization Performance in a Competitive Business Environment." Economic Insights-Trends & Challenges 64.1 (2012).

[7] Ansoff, H.I. Managing Strategic Surprise by Response to Weak Signals. Calif. Manag. Rev. 1975, 18, 21–33. https://doi.org/10.2307/41164635.

[8] Sewlal, R. (2004). Effectiveness of the Web as a competitive intelligence tool. South African Journal of Information Management, 6(1), 1-16. DOI: 10.4102/sajim.v6i1.293.

[9] Sauter, Vicki L. "Competitive intelligence systems: Qualitative DSS for strategic decision making." ACM SIGMIS Database: the DATABASE for Advances in Information Systems 36.2 (2005): 43-57. https://doi.org/10.1145/1066149.1066154.

[10] D. Thorleuchter, T. Scheja, and D. Van den Poel, "Semantic weak signal tracing," Expert Systems With Applications, vol. 41, no. 11, pp. 5009–5016, 2014. DOI:10.1016/j.eswa.2014.02.046.

[11] Herring, J. P. (1999). Key intelligence topics: a process to identify and define intelligence needs. Competitive Intelligence Review: Published in Cooperation with the Society of Competitive Intelligence Professionals, 10(2), 4-14. https://doi.org/10.1002/(SICI)1520-6386(199932)10:2%3C4::AID-CIR3%3E3.0.CO;2-C.

[12] El Akrouchi, M., Benbrahim, H., & Kassou, I. (2020). Early warning signs detection in competitive intelligence. In Proceedings of the 25th International Business Information Management Association Conference—Innovation Vision (pp. 1014-1024).

[13] Ma, Junhong. "Improved K-Means Algorithm in Text Semantic Clustering." (2014). DOI: 10.2174/1874110X01408010530.

[14] Wolter K. (2011) Competitive Intelligence. In: Keuper F., Oecking C., Degenhardt A. (eds) Application Management. Gabler.

[15] Hall, C., & Bensousson, B. (2007). Staying ahead of the competition: How firms really manage their competitive intelligence and knowledge; evidence from a decade of rapid change. New York: World Scientific Publishing Data. https://doi.org/10.1142/6669.

[16] Charity, A. E., & Joseph, I. U. (2013). Manage competitive intelligence for strategic advantage. European Journal of Business and Management, 5(3), 1-9.

[17] Gurel E, Tat M. SWOT analysis: a theoretical review. J Int Soc Res. 2017;1051:994–1006. https://doi.org/10.17719/jisr.2017.1832.

[18] Singh, S.S. (2013). Environment & PEST Analysis : An Approach to External Business Environment, International Journal of Modern Social Sciences.

[19] Veugelers, M.; Bury, J.; Viaene, S. Linking technology intelligence to open innovation. Technol. Forecast.Soc. Chang. 2010, 77, 335–343. https://doi.org/10.1016/j.techfore.2009.09.003.

[20] Holopainen, M., & Toivonen, M. (2012). Weak signals: Ansoff today. Futures, 44(3), 198–205. https://doi.org/10.1016/j.futures.2011.10.002.

[21] Keping, W. (2009). Research on the Enterprise Crisis Early Warning System Based on Competitive Intelligence [J]. Information Studies: Theory & Application, 12.

[22] Zaharia, M.; Xin, R.S.; Wendell, P.; Das, T.; Armbrust, M.; Dave, A.; Meng, X.; Rosen, J.; Venkataraman, S.;Franklin, M.J.; et al. Apache Spark: A Unified Engine for Big Data Processing. Comm. ACM 2016, 59, 56–65. DOI: 10.1145/2934664.

[23] A. Alexopoulos, G. Drakopoulos, A. Kanavos, Ph. Mylonas, G. Vonitsanos, "Two-Step Classification with SVD Preprocessing of Distributed Massive Datasets in Apache Spark", Algorithms, MDPI, March 2020. https://doi.org/10.3390/a13030071.

[24] B. Adil, F. Abdelhadi, B. Mohamed and H. Haytam, "A Spark Based Big Data Analytics Framework for Competitive Intelligence," 2019 1st International Conference on Smart Systems and Data Science (ICSSD), Rabat, Morocco, 2019, pp. 1-6. DOI: 10.1109/ICSSD47982.2019.9002837.

[25] Salloum, S., Dautov, R., Chen, X. et al. Big data analytics on Apache Spark. Int J Data Sci Anal 1, 145–164 (2016). https://doi.org/10.1007/s41060-016-0027-9.

[26] Mühlroth C, Grottke M (2018) A systematic literature review of mining weak signals and trends for corporate foresight. Journal of Business Economics 88(5):643–687.

[27] Yoon, J. Detecting Weak Signals for long-term business opportunities using text mining on Web news. Expert Syst. Appl. 2012, 39, 12543–12550. DOI: 10.1016/j.eswa.2012.04.059.

[28] El Haddadi, A., Dousset, B., & Berrada, I. (2012). Discovering Patterns in Order to Detect Weak Signals and Define New Strategies. In P. Kumar, P. Krishna, & S. Raju (Eds.), Pattern Discovery Using Sequence Data Mining: Applications and Studies (pp. 195-211). Hershey, PA: IGI Global. doi:10.4018/978-1-61350-056-9.ch012.

[29] Thorleuchter, D., & Van den Poel, D. (2015). Idea mining for web-based weak signal detection. FUTURES, 66, 25–34. DOI: 10.1016/j.futures.2014.12.007.

[30] MOREIRA, A. L. M. ; HAYASHI, T. W. N. ; COELHO, G. P. ; SILVA, A. E. A. . A Clustering Method for Weak Signals to Support Anticipative Intelligence. International Journal of Artificial Intelligence and Expert Systems, v. 6, p. 1-14, 2015.

[31] Julien Maitre, Michel Menard, Guillaume Chiron, Alain Bouju, "Détection de signaux faibles dans des masses de données faiblement structurées", Recherche d'information, document et web sémantique, ,vol 3, no.1. doi:10.21494/ISTE.OP.2020.0463.

[32] Gionis, A., Mannila, H., and Tsaparas, P. 2005. Clustering aggregation. In Proceedings of the 21st International Conference on Data Engineering (ICDE) (Tokyo, Japan). https://doi.org/10.1109/ICDE.2005.34.

[33] Youngjung Geum, Jeonghwan Jeon & Hyeonju Seol (2013) Identifying technological opportunities using the novelty detection technique: a case of laser technology in semiconductor manufacturing, Technology Analysis & Strategic Management, 25:1, 1-22, DOI: 10.1080/09537325.2012.748892.

[34] Garcia-Nunes, P.I., & Silva, A.E. (2019). Using a conceptual system for weak signals classification to detect threats and opportunities from web, Futures 107(March 2019):1-16. 10.1016/j.futures.2018.11.004.

[35] Sahbi Sidhom, Philippe Lambert. "Information Design" for "Weak Signal" detection and processing in Economic Intelligence: case study on Health resources. 4th International Conference on Information Systems and Economic Intelligence - SIIE'2011, IGA Maroc, Feb 2011, Marrakech, Morocco. pp.315-321.

[36] Xianjin, Z., & Minghong, C. (2010). Study on early warning of competitive technical intelligence based on the patent map. Journal of Computers, 5(2), 274-281. DOI: 10.4304/jcp.5.2.274-281.

[37] D. M. Thomas and S. Mathur, "Data Analysis by Web Scraping using Python," 2019 3rd International conference on Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, India, 2019, pp. 450-454. DOI: 10.1109/ICECA.2019.8822022.

[38] Patil, C., Baidari, I. Estimating the Optimal Number of Clusters k in a Dataset Using Data Depth. Data Sci. Eng. 4, 132–140 (2019). https://doi.org/10.1007/s41019-019-0091-y.

[39] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. J. Mach. Learn. Res. 3, null (March 2003), 993–1022.

[40] Thomas K Landauer, Peter W. Foltz & Darrell Laham (1998) An introduction to latent semantic analysis, Discourse Processes, 25:2-3, 259-284, DOI: 10.1080/01638539809545028.

[41] Mikolov, Tomas et al. "Efficient Estimation of Word Representations in Vector Space." CoRR abs/1301.3781 (2013): n. pag.

[42] Burkart, Nadia, and Marco F. Huber. "A survey on the explainability of supervised machine learning." Journal of Artificial Intelligence Research 70 (2021): 245-317. DOI:10.1613/jair.1.12228.

[43] Chen, Liang, et al. "A deep learning based method for extracting semantic information from patent documents." Scientometrics 125.1 (2020): 289-312. https://doi.org/10.1007/s11192-020-03634-y.

[44] Wang, Yunli, René Richard, and Daniel McDonald. "Competitive Analysis with Graph Embedding on Patent Networks." 2020 IEEE 22nd Conference on Business Informatics (CBI). Vol. 1. IEEE, 2020. DOI: 10.1109/CBI49978.2020.00009.

[45] Goyal, Palash, and Emilio Ferrara. "Graph embedding techniques, applications, and performance: A survey." Knowledge-Based Systems 151 (2018): 78-94. DOI:10.1016/j.knosys.2018.03.022.

[46] Xiao, W. K. F. X. Z., & Xinyan, L. (2011). Enterprise Competitive Intelligence Crisis Early Warning Review [J]. Journal of Modern Information, 7. DOI: 10.3969/j.issn.1008-0821.2011.07.042.

[47] M. Mohammed, Shapol et al. "A state-of-the-art survey on semantic similarity for document clustering using GloVe and density-based algorithms." Indonesian Journal of Electrical Engineering and Computer Science 22 (2021): 552-562. DOI:10.11591/IJEECS.V22.I1.PP552-562.