A Clustering Method for Weak Signals to Support Anticipative Intelligence

Antonio Leonardo Martins Moreira *antoniolmmoreira@gmail.com*
School of Technology (FT)
University of Campinas (Unicamp)
Limeira, 13484-332, Brazil

Thomas Wiliam Norio Hayashi *thomashsh11@gmail.com*
School of Technology (FT)
University of Campinas (Unicamp)
Limeira, 13484-332, Brazil

Guilherme Palermo Coelho *guilherme@ft.unicamp.br*
School of Technology (FT)
University of Campinas (Unicamp)
Limeira, 13484-332, Brazil

Ana Estela Antunes da Silva *aeasilva@ft.unicamp.br*
School of Technology (FT)
University of Campinas (Unicamp)
Limeira, 13484-332, Brazil

Abstract

Organizations need appropriate anticipative information to support their decision making process. Contrarily to some strategic information analyses that help managers to establish patterns using past information, anticipative intelligence is intended to help managers to act based on the analysis of pieces of information that indicate some sort of trend that may become true in the future. One example of this kind of information is known as a weak signal, which is a short text related to a specific domain. In this work, pairs of weak signals, written in Portuguese, are compared to each other so that similarities can be identified and correlated weak signals can be clustered together. The idea is that the analysis of the resulting similar groups may lead to the formulation of a hypothesis that can support the decision making process. The proposed technique consists of two main steps: preprocessing the set of weak signals and clustering. The proposed method was evaluated on a database of bio-energy weak signals. The main innovations of this work are: (i) the application of a computational methodology from the literature for analyzing anticipative information; and (ii) the adaptation of data mining techniques to implement this methodology in a software product.

Keywords: Anticipative Information, Weak Signals, Clustering, Decision Making Process, Text Mining, Similarity Function.

1. INTRODUCTION

Despite having access to information in different forms from several sources, organizations still suffer from the lack of information that can effectively help to anticipate market tendencies and support the process of strategic decision-making [1], [2], [3], [4], [5]. This is a significant issue as these organizations need to detect tendencies as soon as possible, so that they can modify their internal processes and improve their adaptation to deal with uncertainties, threats and opportunities that may arise in near future [6].

Two approaches have been adopted by organizations in order to try to fulfill this need. The first one, known as *prevision*, consists in the accumulation of historical data in order to try to identify tendencies. In this context, the decision-making process is assisted by the application of *Data Mining* (DM) techniques [7] to the large amount of data saved over the years, in an attempt to discover potentially useful patterns that indicate how data are related until that specific moment in time [8]. This analysis, which can be seen as a way to predict future events considering information that is already consolidated and established, can lead to the prediction of events and circumstances that support actions related to business transactions. Therefore, it is possible to call this first approach *Predictive Intelligence*.

The second approach that can be used to help organization managers to get relevant and fresh information is known as *Anticipative Intelligence* (AntInt). AntInt's goal is to capture information that can be useful to identify changes in social-economic areas, thus allowing the anticipation of opportunities and threats [9],[10]. According to [11], although hard to identify and interpret, this kind of information is vital to the success and even to the survival of organizations. Formally, AntInt can be defined as a collective, pro-active and continuous process that uses pertinent information about both the environment and the environmental changes in order to create business opportunities, innovation and to reduce risks [12].

One of the most well known methodologies for capturing and analyzing anticipative information is known as the L.E.SCAnning® Methodology [12]. This approach, which tries to model strategic information in order to make it collectively available, consists of seven steps that implement the use of concepts of AntInt into an organization: (1) *application's domain* step, which corresponds to the analysis of the domain of the problem; (2) *aim specification* step, in which actors and sources related to the domain of the application are specified by experts in the field; (3) *acquisition*, which is the process of collecting anticipative information about actors and sources defined in (2); (4) *information selection* and *collective presentation*, which corresponds to the analysis of the anticipative information and presentation of the results to the organization; (5) *memory step*, which consists in storing anticipative information; (6) *collective sense creation* step, in which different information is analyzed by several people in order to create a resulting hypothesis; and (7) *hypothesis spreading* step, in which the results of the information analyses are shared in order to build the *collective knowledge*.

Although such methodologies for capturing and analyzing anticipative information provide the general steps for implementation in organizations, it is known that they still lack from computational tools that can support their steps [10], [13]. Therefore, this work intends to provide one step in the direction of providing computational tools capable of supporting AntInt methodologies, more specifically the steps of analysis of anticipative information in the L.E.SCAnning® Methodology (numbered, (4), (5) and (6) in the previous discussion).

In order to analyze anticipative information, it is proposed here the use of algorithms originally developed to perform *clustering*, which is an unsupervised data mining approach that intends to group a set of *patterns* (usually vectors in a multidimensional space) into *clusters* according to the similarity of those patterns [7]. The rationale behind the application of clustering algorithms to analyze anticipative information is that if a given anticipative information repeats itself in different pieces of data or if it is reinforced by similar ones, then it is possible that this group of information may lead to the formulation of a hypothesis that results in strategic actions in an organization [14]. Therefore, the idea here is to group similar information obtained by different sources, in order to ease the prediction of changes (change of events, emergence of new technologies etc.) that could lead the market to a new business direction.

This work will deal with pieces of anticipative information known as *weak signals*, which are short sentences that can be captured from several kinds of sources (news, advertisements, interviews etc.). These weak signals can be seen as disperse pieces of information that, in isolation, do not seem to be relevant. However, when combined with other similar weak signals, they may induce

a new perception about the subject, which can lead to anticipative decisions [9], [10]. It is important to point out that, in this work, weak signals in Portuguese will be considered.

In order to automatically group weak signals from different sources, the *K-medoids* algorithm [15], which is a *partitioning clustering* algorithm (it decomposes the dataset into a set of disjoint clusters) based on the well known and frequently used *K-means* [16], was adopted here. Although weak signals are basically short texts, the clustering process of weak signals presents distinct properties when compared to clustering of whole documents and texts [17]. For example, texts are generally clustered first by defining a set of selected terms that are representative of the field, and then by evaluating the similarity of texts according to the occurrence of those terms in each text. Such approach does not apply to clustering of weak signals, as, given that the goal here is to identify new information, it is not possible to previously limit the set of representative terms. Therefore, the similarity between two weak signals must be evaluated differently. All these aspects were considered in this work, and will be further discussed in Section 3.

This paper is organized as follows. In Section 2, the main concepts about weak signals will be discussed, together with a brief presentation of the main tools from the literature for monitoring and analyzing weak signals. The proposal of this work will be thoroughly presented in Section 3, while the results obtained by the application of the proposed technique to a database of bioenergy-related weak signals will be detailed and discussed in Section 4. Finally, some conclusions and indications for future work will be provided in Section 6.

2. WEAK SIGNALS: DEFINITIONS, TOOLS AND METHODS

Weak signals [9], [10], [18] are disperse elements of information that are not always perceived as important when considered in isolation, but that can reveal certain ideas or induce new perceptions that may lead to anticipative decisions when analyzed together with other similar weak signals. According to [19], weak signals are incomplete, fragmented and meaningless when analyzed individually, but can make sense if seen as a set of information. Weak signals can be captured from a multitude of sources, such as an advertisement, a picture, an image and a drawing, among others.

According to [20], which was the first to propose the concept of weak signals, the preoccupation with the organization's dynamics and technologies leads to an inability to rapidly cope with threats and opportunities that arise. In this case, weak signals introduce a new way to look at problems: they neither express concern about a specific problem nor about the use of a specific technology. As a result of real world observations made by specialists in the field, weak signals bring *fragments of possibilities* that, together, can lead to the construction of different scenarios.

In order to illustrate how a set of weak signals may lead to the formulation of hypotheses that indicate possible future steps that can be adopted by an organization, Figure 1 presents an example of a set of similar weak signals obtained from newspaper headlines and one possible hypothesis that could be formulated from the analysis of these weak signals. The headlines in Figure 1 were translated from Portuguese and a fictional label ("ACME") replaced the name of the original company.

The process of organizing weak signals and formulating hypothesis as illustrated in Figure 1 is still considered ill structured and inadequately understood [10], [21]. One of the reasons for that is that there are only a few tools for organizing and analyzing anticipative information. Some of these tools will be briefly discussed in what follows.

Group of similar weak signals:

1) Brazilian market is one of ACME's priorities.
2) ACME will contribute to a social organization of the Brazilian government.
3) ACME will build new plants in Brazil.

Hypothesis formulated by ACME's rival:
Support the Brazilian government in social plans in order to get advantage over ACME.

FIGURE 1: Example of hypothesis creation from a set of weak signals.

In [22], it is proposed a method for displaying and analyzing sentences, which allows the monitoring of anticipative information in different organizational environments: political, economic, technological, social and related to competition. Although the proposed method follows some theoretical steps, it does not provide any computational algorithm to perform the information analysis. Specialists from each area are responsible for such task.

In [23] a tool named "*Puzzle*" was proposed for the analysis of weak signals. The general idea behind Puzzle is the creation of a *map* or a *puzzle* that tries to visually organize the weak signals (which are considered pieces of the puzzle). Such visual organization is made with the goal of helping managers to analyze the weak signals and formulate hypothesis. Although Puzzle does help managers to formulate hypothesis, there are still some uncovered points: the analysis of bases of weak signals with several sentences can be a time consuming and error prone activity.

The use of a clustering technique implemented in the software package named CLUTO [24] to group unstructured texts obtained from an exploratory research in the Internet was proposed in [25]. In [25], potential weak signals are automatically detected in those text documents and such documents are clustered according to the similarity of weak signals contained in each of them. After that, experts analyze the clusters and the weak signals obtained.

Although this proposal [25] is also based on clustering, it contains fundamental differences when compared to what is being proposed here. The main difference lies in the way the weak signals are extracted. As the proposal of this work intends to develop a specific tool to support some steps of the L.E.SCAnning® Methodology, here experts are supposed to previously identify, extract and store the weak signals, and then a clustering algorithm is applied to such dataset. The rationale of this approach is based in two aspects: (*i*) the L.E.SCAnning® Methodology clearly separates the steps of capture and storage from the steps of analysis of weak signals; and (*ii*) it is not trivial for an expert to identify weak signals, so it is possible to say that it is even harder to obtain algorithms capable of obtaining satisfactory results in this particular task.

Another difference between the technique proposed here and the one adopted in [25] is that the latter adopts the K-means algorithm, while here the K-medoids algorithm was implemented and tested. And, finally, in this work a base of synonyms is also employed to support the clustering process.

Therefore, in summary it is possible to say that, when compared to previous approaches from the literature, the technique proposed here presents the following advantages:

- It preserves the human-based capturing of weak signals.
- It previously clusters the base of weak signals, which reduces the amount of information to be dealt with, thus facilitating the analysis of such information. In this approach, weak signals are presented in groups to the user, and not individually.

- As the clustering step is performed independently and apart from the extraction of the weak signals, it is possible to adjust or even replace the clustering module according to the characteristics of the weak signals database.

In the next section, the technique proposed here to cluster bases of weak signals will be presented in details.

3. CLUSTERING WEAK SIGNALS WITH K-MEDOIDS

The fundamental aspect of any type of clustering or grouping process is the definition of how each item in the dataset is related to the others. Specifically in the context of this work, it is possible to identify three main types of relations between weak signals: *cause-effect*, *similarity* and *contrast* [10], [26]. As the goal here is to group similar weak signals to ease the hypothesis formulation step for the user, the focus will be in *similarity* relationships, while the study of cause-effect and contrast will be left for future work. It is possible to define several similarity metrics between pairs of weak signals, according to the final goal of the information retrieval system being developed. In this work, similarity will be evaluated according to the *number of similar words* observed in two weak signals, being considered as similar words *identical* words, words with the same *stem* and even *synonyms*.

Besides the identification of the type of relation among weak signals that must be considered by the clustering algorithm, it is mandatory to perform some *preprocessing* of the weak signals before clustering, which is known to be one of the most important steps in pattern recognition of texts [27]. In this work, three activities constitute the preprocessing step: *elimination of stop-words*, *stemming* and *search for synonyms*.

All details about the preprocessing step, similarity metric and clustering algorithm adopted in this work will be discussed in the following sections.

3.1 Preprocessing and Storing the Weak Signals

In traditional works of text mining, it is common to have a set of previously selected terms that composes a representative group of terms of the field [27]. Once these terms have been selected, it is necessary to structure all the documents to be mined. In order to do so, one way of structuring documents is to convert each document into a *feature vector*, which basically corresponds to a vector in a multi-dimensional space, being the *value* in each dimension associated with one of the representative terms previously selected. Formally, in this approach, called *bag-of-words*, each document i is represented as a vector $\mathbf{d_i} = (s_{i1}, s_{i2}, s_{i3}, ..., s_{im})$, where s_{ij}, j=1…m, is a value that relates document i to term j. The value s_{ij} can be obtained according to different approaches: for example, s_{ij} can simply indicate whether a specific term j is present in document i, or it can indicate the *frequency* of occurrence of term j in document i [28].

Although bag-of-words is the main approach for text mining in the literature, it was not adopted in this work due to the type of text that must be analyzed (weak signals). The sentences considered here represent anticipative information, so it is not possible to previously limit the analysis to a specific subset of terms, as such approach would compromise the goal of finding tendencies and relations among the weak signals in a broad set of contexts. Even if the compromise of the innovation aspect of anticipative information analysis is not considered, the adoption of bag-of-words would still present problems, as the feature vector of each weak signal would be *sparse* (a vector with most of the values null), since weak signals are often short sentences.

Due to the exposed factors, feature vectors were not adopted here. In contrast, the clustering process is performed according to a *similarity matrix* with structure equivalent to the one presented in Table 1. Such similarity matrix must be obtained and stored *before* the clustering step.

	WS_1	**WS_2**	**...**	**WS_m**
WS_1	v_{11}	v_{12}	…	v_{1m}
WS_2	v_{21}	v_{22}	…	v_{2m}
...	…	…	…	…
WS_m	v_{m1}	v_{m2}	…	v_{mm}

TABLE 1: Structure of the similarity matrix adopted in the clustering process.

The matrix illustrated in Table 1 stores the values of *similarity* v_{ij} between each pair of weak signals ws_i and ws_j. Such values of similarity v_{ij} are given by a combination of the number of identical words, synonyms and different words observed in the pair of weak signals *i* and *j*. Further details about the *similarity metric* adopted here will be given in Section 3.2.

As the clustering algorithm adopted here can group sets of weak signals according to similarity matrices such as the one illustrated in Table 1, there is no need to either predefine a set of representative terms or build feature vectors for each weak signal. However, in order to obtain such similarity matrix, a series of preprocessing steps must be made, as it will be explained in the following subsections.

3.1.1 Elimination of stop-words

The first preprocessing step that must be performed to the base of weak signals is the *elimination of stop-words*, which are repetitive words that do not add any meaning to the weak signal. In order to eliminate stop-words, all words in a weak signal must be compared to those stored in a previously defined *stop-list* (a list containing all stop-words for a given language).

The main goal of this preprocessing step is to *reduce* the dimension of the problem, as the number of terms in each weak signal is invariably lowered, thus reducing the computational cost required to deal with such weak signal [29]. In this work, a Portuguese list of stop-words was used. This list was based on the work of [30]. The list contains: articles, indefinite pronouns, definite pronouns, prepositions and adverbs.

3.1.2 Stemming

After stop-words are removed from the weak signals, the next step consists in the identification of *morphological variations* of each term, in a process called *stemming*. Stemming, which has been widely used in information retrieval from texts [31], [32], is a computational process that reduces all possible variation of a word to a unique form: its *stem* [33].

As the stem of a given word corresponds to an unique structure that represents all its variations, the use of *stemmers* in the base of weak signals tend to reduce redundancy and approximate weak signals that contain syntactically similar words. Therefore, it is possible to say that stemming has an impact in the relation of similarity among weak signals.

Given that the focus of this work is not the development of a new stemmer for Portuguese language, the proposal of [32] named *Suffix Extractor for Portuguese Language* (from Portuguese, *Redutor de Sufixos para Língua Portuguesa* – RSLP), was adopted here. The stemmer of [32], which was later improved by [34], is capable of dealing with irregular verbs, proper nouns, all the plural rules found in Portuguese and also of treating exceptions. RSLP was applied to all weak signals in the database considered in this work.

3.1.3 Search for Synonyms

The final preprocessing step performed to the base of weak signals is the search for synonyms, in order to allow the identification of semantically related concepts that may be present in a given pair of weak signals. The use of synonyms to identify semantic relationships has been previously applied in some works from the literature [35], as this approach, when compared to traditional methods based on the frequency of terms, tends to allow a better evaluation of whether two texts relate to the same subject even when different words are used.

In this work, a lexical base called TeP 2.0 [36], which is an electronic thesaurus for Brazilian Portuguese that contains almost twenty thousand groups of synonyms, was used to identify the synonyms of the stemmed words that belong to the weak signals. The number of synonyms identified for each stemmed word in a weak signal is used in the calculation of the *similarity function* adopted in the clustering process, as will be discussed in the following section.

3.2 Clustering Algorithm and Similarity Function

After preprocessing the set of weak signals, the second step of the proposed technique is to apply a clustering algorithm to divide the weak signals into subsets according to their similarity.

To perform clustering, the algorithm called *K-medoids* was adopted in this work, more specifically the variation of *K-medoids* known as *Partitioning Around Medoids* (PAM) [15]. To evaluate the similarity among weak signals, a similarity function was proposed here, based on Jaccard's similarity metric [37].

K-medoids is based on the well-known and frequently used *K-means* algorithm [16]. Both *K-means* and *K-medoids* are *partitioning clustering* algorithms, which means that they decompose the dataset into a set of disjoint clusters. Both algorithms are also based on the representation of the clusters by one element that corresponds to the *center* of the cluster. These *central elements* are called *centroids* (in *K-means*) or *medoids* (in *K-medoids*).

In *K-means*, the centroid of each cluster is defined by the *mean* (usually weighted average) of the feature vectors of all the samples that belong to that cluster. Therefore, *K-means* works conveniently with datasets that allow the representation of each sample as a feature vector with numerical components. However, this also implies that outliers can negatively affect the results of *K-means*.

In *K-medoids*, on the other hand, the center of each cluster (or *medoid*) is represented by one of the data samples that belong to the cluster, which is the element with the highest average similarity with respect to all other elements in the cluster. As *K-medoids* only needs the similarity values among the samples of the dataset, it is particularly suitable to those situations in which it is not possible obtain a numerical feature vector for each sample.

Therefore, the choice of *K-medoids* to be applied in this work was made based on two main aspects [38]: (*i*) as discussed in Section 3.1, it is not possible to represent each weak signal by a feature vector, so *K-medoids* suits perfectly to this scenario; and (*ii*) the choice of medoids is dictated by the location of the predominant fraction of data samples inside a cluster, which is an approach less sensitive to the presence of outliers.

A variation of the well-known PAM [39] (*Partitioning Around Medoids*) algorithm was adopted in this work. In summary, PAM operates on a similarity matrix of a given data set and maximizes a sum of similarities between objects of the data set. Each cluster is represented by one of the objects in the cluster. This object is called the medoid. A medoid can be defined as that object of a cluster, whose average similarity to all the objects in the cluster is maximal. After finding the set of medoids, each object of the data set is assigned to the nearest medoid.

In this work, the implementation of the clustering algorithm is the classical PAM algorithm, in which the first set of medoids is randomly chosen. As *K-medoids* is capable of clustering datasets that are not represented by feature vectors, the similarity matrix presented in Section 3.1 has to be obtained before the application of the algorithm. In this work, an adaptation of the Jaccard similarity metric was proposed, in order to evaluate how similar each pair of weak signals are. The original Jaccard similarity metric, given in Eq. (1), is widely adopted in pattern recognition [37].

$$Jaccard(o_j, o_i) = \frac{f_{11}}{f_{01} + f_{10} + f_{11}}, \quad (1)$$

where f_{11} is the number of terms that are present in both documents o_i and o_j, f_{01} is the number of terms that are missing in o_j but are present in o_i and f_{10} is the number of terms that are present in o_j and missing in o_i.

The adaptation of Jaccard's metric to evaluate the similarity between pairs of weak signals considers the numbers of *equal words*, *synonyms*, and *different words* between two weak signals, as given in Eq. (2):

$$similarity(o_j, o_i) = \frac{numSyn(o_j, o_i) + numEqual(o_j, o_i)}{numSyn(o_j, o_i) + numEqual(o_j, o_i) + numDiff(o_j, o_i)} \quad (2)$$

where $numSyn(o_j, o_i)$ is the number of synonyms that belong to weak signals o_j and o_i, $numEqual(o_j, o_i)$ is the number of identical words between weak signals o_j and o_i, and $numDiff(o_j, o_i)$ is the number of different words between weak signals o_j and o_i.

Given that the proposed approach is supposed to be applied to weak signal databases that contains *n* sentences that must be grouped into *k* disjoint groups (*k* defined by the user), being each group represented by a set of elements which belong to *Total_of_Elements* ={o_1, o_2, ..., o_n} and each weak signal by o_i, i=1...n, the general steps of the algorithm can be summarized as given in Figure 2.

1. Arbitrarily choose the number *k* of groups that should be extracted;
2. For each weak signal, eliminate all stop-words;
3. Using the RLSP stemmer, stem each remaining words for each weak signal;
4. For each stemmed word, find its synonyms in the TeP 2.0 lexical base;
5. Calculate similarity between each pair of weak signals (o_j, o_i) using the function given by (2);
6. Choose the initial *k* medoids randomly, one for each group;
7. Repeat
   7.1. Assign each weak signal to its nearest medoid using the similarity metric calculated in step 5;
   7.2. For each group C_j, recalculate its medoid according to:

$$medoid_j = max\{\, o_i \in C_j,\ i = 1,\ldots, n;\ \sum_{l=1}^{n} |similarity(o_l, o_i)|/n \,\},$$

   where C_j is the j_{th} cluster of *k* elements, o_i is the chosen medoid, o_j represents all other objects in the data base, whose similarities are calculated to o_i and *similarity*(o_j, o_i) was calculated in step 5;

FIGURE 2: Example of hypothesis creation from a set of weak signals.

4. RESULTS AND DISCUSSION

The variation of the *K-medoids* algorithm (see Section 3.2) was evaluated here on a database of 118 weak signals from the domain of bioenergy. All weak signals in this database are in Portuguese and were identified and collected by experts in the field.

The algorithm was tested in the scenario in which k=4 clusters had to be obtained for the dataset, and they were ran for 50, 100 and 1000 iterations each. With 1000 iterations the algorithm converged, i.e., no more changes in the structure of the clusters were observed in the final iterations.

In order to evaluate the quality of the partitions obtained for the bio-energy dataset, two metrics were considered: average intra- and inter-group similarities. Intra-group similarity is calculated by the mean of the sum of all similarity values of the cluster elements (given by the weak signal in a cluster and the medoid of the cluster) and the number of elements in this cluster. Inter-group similarity is given by the average of the similarities of each weak signal in a cluster to each weak signal assigned to remaining clusters. Given these two metrics, the clustering process is supposed to implicitly maximize intra-group similarity, while simultaneously minimizing the inter-group similarity.

The experimental intra-group and inter-group similarities are shown in Table 2. Table 2 presents the results obtained by the application of the *K-medoids* algorithm as given in Figure 2 combined with the similarity metric given in Eq. 2.

Evaluation Metric	Number of Iterations	K-medoids
Intra-group similarity	50	0.207361
Intra-group similarity	100	0.209142
Intra-group similarity	1000	0.209978
Inter-group similarity	50	0.180599
Inter-group similarity	100	0.176939
Inter-group similarity	1000	0.176800

TABLE 2: Intra- and Inter-group similarities obtained for the *K-medoids* algorithm, for 50, 100 and 1000 iterations, using the similarity function defined in Eq. 2.

Considering the results shown in Table 2, it is possible to notice that intra-group similarities increased with the number of iterations. With respect to inter-group similarities, as expected the values of this metric decreased with the increase of the number of iterations.

Although the results shown in Table 2 show that, with respect to clustering a dataset, the proposed technique behaves as expected, it is important to emphasize that the final goal is to obtain a solution that simplifies the decision making process. The method would present to the decision maker four groups of weak signals, and the analysis of these groups should help him/her to formulate a hypothesis that would contribute to future strategic actions.

In order to illustrate the clustering analysis process, Table 3 presents the five most similar weak signals - in relation to the medoid - in a given group.

They are presented from the most similar weak signal to the least similar weak signal in relation to the group's medoid. This group was chosen because it presented the best intra-group similarity. It is important to mention again that the medoid is also a weak signal, particularly the one with the highest similarity to all other weak signals in the cluster.

Weak signal one is the most similar to the medoid in the chosen group. Similar words between the two sentences are: solar, energy and plant. Synonyms are: build and construction. It is important to emphasize that the clustering algorithm also stems words that are synonyms. Therefore, a word like construction is steamed to construct. After the steaming process, the word construct is established as a synonym of build which is already in its steam. It is also important to remember that all sentences are in the Portuguese Language.

Medoid	President of X believes that photovoltaic solar energy will have its peak in Brazil in two years and he has plans to build a plant in 2012.
Weak signal 1	It is in advanced phase the construction of a solar energy plant in CEARA.
Weak signal 2	The XYZ Energy industry opens its first plant of biodiesel.
Weak signal 3	Wind energy may be a development factor in Brazil.
Weak signal 4	Spain wants to invest in solar energy in Piauí.
Weak signal 5	Institutes are discussing the directions of solar energy in Brazil.

TABLE 3: Medoid and the five most similar weak signals (with respect to the medoid) from one of the clusters.

The analysis of weak signals presented in Table 3 indicates some important strategic points that could be used to formulate hypothesis of actions:

- Clean energy is the main promise for energy projects and partnerships;
- Brazil is a promising country to invest in clean energy;
- There will be a reasonable investment in solar energy in Brazil;
- Brazil is trying to create a forum for discussing the directions of the creation of plants of solar energy.

Therefore, if a country, such as Spain, as mentioned in weak signal 4, desires to invest in clean, such as solar energy, in Brazil, it would be advisable to observe: the location and the type of clean energy. Another hypothesis that could be raised from the analysis of the weak signals is that locations where solar energy plants are being developed are candidates to become promising regions to invest.

In order to evaluate the opposite results Table 4 presents the five least similar weak signals - in relation to the medoid - of the same cluster given in Table 3. They are presented from the least similar to the most similar weak signal (compared to the medoid). It is possible to observe that, the weak signals that are less similar to the medoid tend to have more words than the ones presented in Table 3. This is due to the fact that the more different words there are between the weak signal and the medoid the least is the similarity between them. In this way, longer sentences are penalized, as it is likely that they have more different words to the medoid than a shorter sentence. Another point to be considered in Table 4 is that, although presenting the least values of the similarity function to the medoid, weak signals one to five maybe be important to be analyzed in the decision making process.

The information, as described, for example, by weak signal four, maybe an announcement of more use of ethanol as fuel in Congo what reinforces the development of clean energy. The problem is that the clustering algorithm did not find any similar words and found few synonyms between weak signal four and the medoid. However, weak signal four is in the same cluster of the weak signals of Table 3 (to which it is related even though with low similarity by the algorithm) and, consequently, it shares some similarity to the medoid.

This observation points out to the need for a base of synonyms not only with words from the Portuguese Language, but a base of synonyms with regard to the domain problem, in this case, bio-energy domain. This would be more effective to find weak signals that are similar regarding their domain. Considering again the example of weak signal four, if there was a domain-oriented base of words, the word "ethanol", for example, would be considered as an important one as a source of clean energy. This fact could be taken into consideration when comparing this weak signal to the medoid that contains "solar energy" which is also a source of clean energy.

Medoid	President of X believes that photovoltaic solar energy will have its peak in Brazil in two years and he has plans to build a plant in 2012.
Weak signal 1	WG, which builds of electronics and motors, will buy an unit of the EX company.
Weak signal 2	The Air XC air company completed today its first test with an airplane using 50% of clean fuel and 50% of traditional fuel.
Weak signal 3	Researchers of Pará and São Paulo will concentrate efforts to produce enzimatic compounds to degrade cellulose in order to enable ethanol from cellulose.
Weak signal 4	The WZ company enables technicians from Congo in producing sugar cane.
Weak signal 5	The XS industry will invest in projects to dominate the process of silicon enrichment, which is the last phase that Brazil needs to fully dominate the technology to produce photovoltaic energy.

TABLE 4: Medoid and the five least similar weak signals (with respect to the medoid) from the same cluster presented in Table 3.

In order to orient decision maker actions, step 6 of the L.E.SCAnning® Methodology [9],[10],[12] could be applied to the group presented in Table 3. Step 6 concerns the collective sense creation. In order to create sense, some questions could be asked regarding the weak signals, in order to create relations between them, such as:

- Is the weak signal an isolate case?
- Is there any contradiction among the weak signals?
- Which are the differences and similarities among weak signals?

As mentioned in section 2, another work that supports the creation of sense in weak signals is proposed by [23]. This work suggests a way of creating sense from weak signals by making a puzzle in which weak signals are pieces of the puzzle. The focus of the work is to create new ideas from existing ideas. The innovation of the work is a prototype tool to create the relations. In order to create the puzzle, [23] suggests the connection of weak signals using cause-effect, similarity and contradiction relations.
The current work could contribute with groups, such as Table 3, for example, which would be the starting point to the creation of these relations, as a group contains similar weak signals and this can facilitate the creation of relations.

The creation of sense could lead the decision make team to the formulation of an action plan, i.e., a hypothesis that could feed the decision making process. Current work could support the creation of sense with groups of similar weak signals.

5. CONCLUSIONS AND FUTURE WORKS

This work proposed the creation of an automatic method for clustering anticipative information. This kind of information is different from classical strategic information, as it shows tendencies and views of possible events in future.

Numerical results showed that it is possible to group weak signals using *K-medoids* and a similarity function composed by the parameters: number of equal words, number of synonyms and number of different words.

Considering factors such as the need for anticipative information in organizations and the need to automate weak signals clustering and analysis, it is possible to say that this work brings as innovations:

- Implementation of parts of a methodology for clustering and analyzing anticipative information;

- Adaptation of text mining techniques to implement this methodology in a software product;
- Implementation and test of the k-medoids clustering algorithm to anticipative information;
- Treatment of anticipative information written in the Portuguese Language.

However, there is still some work to be done. As mentioned previously, weak signals are a different form of information that demands different approaches of analysis. From this work it is possible to affirm that there is a possibility of investing in *K-medoids* and similarity functions, such as the one used in this work, as a good possibility to cluster weak signals.

As future works, the authors intend to extend the proposed methodology to the creation of:

- A domain-oriented base of synonyms;
- The application of different similarity functions;
- Output reports showing parameters such as:
  - Number of synonyms between weak signals of a specific group;
  - Number of equal words between weak signals of a specific group;
- A base of hypothesis resulting from the analysis of groups.

6. ACKNOWLEDGEMENTS

The authors would like to thank FAPESP for the financial support (process numbers: 2011/08696-9 and 2012/11917-0).

7. REFERENCES

[1] S. Haeckel. “Peripheral vision: Sensing and acting on weak signals making meaning out of apparent noise: The need for a new managerial framework”. Long Range Planning, vol. 37, pp. 181-189, 2004.

[2] Z. Xianjin and Y. Feng. “Research on the Acquirement of Enterprise Risk Competitive Intelligence Based on Data Mining”, in Proc. of the 5th International Conference on Wireless Communications, Networking and Mobile Computing (Wicom'09), 2009, pp. 1-4.

[3] C. Ting; N. Xiao and Y. Weiping. “The Application of Web Data Mining Technique in Competitive Intelligence System of Enterprise based on XML”, in Proc. of the 3rd International Symposium on Intelligent Information Technology Application, 2009, pp. 396-399.

[4] M.-J. Shih; D.-R. Liu and M.-L. Hsu. “Discovering competitive intelligence by mining changes in patent trends”. Expert Systems with Applications, vol. 37(4), pp. 2882–2890, 2010.

[5] K. Xu; S. S. Liao; J. Li and Y. Song. “Mining comparative opinions from customer reviews for Competitive Intelligence”. Decision Support Systems, vol. 50(4), pp. 743–754, 2011.

[6] P. Rossel. “Weak signals as a flexible framing space for enhanced management and decision making”. Technology Analysis & Strategic Management, vol. 21(3), pp. 307-320, 2009.

[7] J. Han and M. Kamber. Data Mining: Concepts and Techniques (2nd edition). Walthan, MA: Morgan Kaufmann, 2006.

[8] S. Delisle. “Towards a Better Integration of Data Mining and Decision Support via Computational Intelligence”, in Proc. of the 16th International Workshop on Database and Expert Systems Applications, 2005, pp. 720-724.

[9] R. Janissek-Muniz; H. Lesca and H. Freitas. “Inteligência Estratégica Antecipativa e Coletiva para Tomada de Decisão”. Revista Organização em Contexto (Jul/Dez, 2007), pp. 92-118.

[10] H. Lesca and N. Lesca. Weak Signals for Strategic Intelligence: Anticipation Tool for Managers**.** Hoboken, NJ: Wiley, 2011.

[11] G.S. Day and P.J. Schoemaker. “Scanning the periphery”. Harvard Business Review*,* vol. 1(12), pp. 135-149, 2005.

[12] H. Lesca. “Veille stratégique: La méthode L.E.SCAnning®”. Editions EMS, 2003 180p.

[13] T. Kuosa. “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*,* vol. 42(1), pp. 42-48, 2010.

[14] S. Mendonça; G. Cardoso and J. Caraça. “Some Notes on the Strategic Strength of Weak Signal Analysis”. Lini Working Papers no. 2. Available: http://www.lini-research.org/np4/working_papers [Jan. 13, 2015].

[15] S.-C. Chu; J.F. Roddick and J.-S. Pan. “Improved search strategies and extensions to *K-medoids*-based clustering algorithms”. International Journal of Business Intelligence and Data Mining, vol. 3(2), pp. 212-231, 2008.

[16] L. Rokach and O. Maimon. “Clustering Methods” in Data Mining and Knowledge Discovery Handbook, 1st ed., O. Maimon and L. Rokach, Eds. New York: Springer, 2005, pp. 321-352.

[17] J. Oliva; J.I. Serrano; M.D. del Castillo and A. Iglesias. “SyMSS: A syntax-based measure for short-text semantic similarity”. Data & Knowledge Engineering, vol. 70(4), pp. 390-405, 2011.

[18] M. Holopainen and M. Toivonen. “Weak signals: Ansoff today”. Futures, vol. 44(3), pp.. 198–205, 2012.

[19] M.-L. Caron-Fasan and R. Janissek-Muniz. "Análise de Informações de Inteligência Estratégica Antecipativa: Proposição de um Método, Caso Aplicado e Experiências". Revista de Administração da Universidade de São Paulo, vol. 39(3), pp 205-219, 2004.

[20] H. I. Ansoff. “Managing Strategic Surprise by Response to Weak Signals”. California Management Review, vol. 18(2), pp. 21, 1975.

[21] J.G. Walls; G.R. Widmeyer and O.A. El Sawy. “Building an Information system design theory for vigilant EIS”. Information System Research, vol. 3(1), pp. 36-59. 1992.

[22] A. Ozaki; A. Del Rey and F.C. Almeida. **“**Radar de Monitoramento Tecnológico: Uma Ferramenta de Interpretação de Sinais Fracos para Identificação de Surpresas Estratégicas”. Future Studies Research Journal, vol. 3(1), pp. 84-110, 2011.

[23] K. Rouibah and S. Ould-al. “PUZZLE: a concept and prototype for linking business intelligence to business strategy”. The Journal of Strategic Information Systems, vol. 11(2), pp. 133-152, 2002.

[24] G. Karypis. “CLUTO: A clustering toolkit”. Technical Report 02-017, College of Science and Engineering, University of Minnesota. Available: http://www.cs.umn.edu/research/technical_reports/view/02-017 [Jan. 14, 2015]

[25] N. Tabatabaei. “Detecting Weak Signals by Internet-Based Environmental Scanning**”.** M.A. thesis, University of Waterloo, Canada, 2011.

[26] R. Janissek-Muniz; H. Lesca and H. Freitas. “Inteligência Estratégica Antecipativa e Coletiva para Tomada de Decisão”. Revista Organizações em Contexto, vol. 2(4), pp. 92-118, 2005.

[27] R. Feldman and J. Sanger. The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data**.** Cambridge, MA: Cambridge University Press, 2006.

[28] G, Salton; J. Allan and A. Singhal. “Automatic text decomposition and structuring”. Information Processing & Management, vol. 32(2), pp. 127–138, 1996.

[29] C.D. Manning; P. Raghavan and H. Schütze. An Introduction to Information Retrieval. Cambridge, MA: Cambridge University Press, 2008.

[30] M.A.L. Dias. “Automatic extraction of Portuguesekey-words applied to dissertations and thesis in the engeneering area”. M.Sc. thesis, University of Campinas (Unicamp), Brazil, 2004.

[31] M.A. Yunus; R. Zainuddin and N. Abdullah. “Visualizing Quran documents results by stemming semantic speech query”, in Proc. of the 2010 International Conference on User Science and Engineering (i-USEr), 2010, pp. 209-2013.

[32] V.M. Orengo and C. Huyck. “A Stemming Algorithm for the Portuguese Language”, in Proc. of the 8th International Symposium on String Processing and Information Retrieval (SPIRE), 2001, pp. 186-193.

[33] J.B. Lovins. “Development of a Stemming Algorithm”. Mechanical Translation and Computational Linguistics, vol. 11(1-2), pp. 22-31, 1968.

[34] A.R. Coelho. **“**Stemming para a língua portuguesa: estudo, análise e melhoria do algoritmo RSLP”. Undergraduate Paper, Federal University of Rio Grande do Sul (UFRGS), Brazil, 2007.

[35] R. Baghel and R. Dhir. “Text Document Clustering Based on Frequent Concepts”, in Proc. of the 1st International Conference on Parallel, Distributed and Grid Computing (PDGC), 2010, pp. 366-371.

[36] E. G. Maziero; T.A.S. Pardo; A. Di Felippo and B. C. Dias-da-Silva. “A base de dados lexical e a interface web do TeP 2.0: thesaurus eletrônico para o Português do Brasil”, in Proc. of the XIV Brazilian Symposium on Multimedia and the Web, 2008, pp. 390-392.

[37] N.A.A. Aziz; S.S. Salleh and D. Mohamad “Investigating Jaccard Distance Similarity Measurement Constriction on Handwritten Pen-based Input Digit”, in Proc. of the 2010 International Conference on Science and Social Research (CSSR), 2010, pp. 1181-1185.

[38] N. Fanizzi; C. d’Amato and F. Esposito. “A Hierarchical Clustering Procedure for Semantically Annotated Resources”, in Proc. of the 10th Congress of the Italian Association for Artificial Intelligence, 2007, pp. 266-277.