A meaningful information extraction system for interactive analysis of documents

Julien Maitre, Michel Ménard, Guillaume Chiron, Alain Bouju, Nicolas Sidère

A meaningful information extraction system for interactive analysis of documents

Julien Maitre, Michel Ménard, Guillaume Chiron, Alain Bouju, Nicolas Sidère
L3i, Faculty of Science and Technology
La Rochelle University
La Rochelle, France
{julien.maitre, michel.menard, guillaume.chiron, alain.bouju, nicolas.sidere}@univ-lr.fr

***Abstract*—This paper is related to a project aiming at discovering weak signals from different streams of information, possibly sent by whistleblowers. The study presented in this paper tackles the particular problem of clustering topics at multi-levels from multiple documents, and then extracting meaningful descriptors, such as weighted lists of words for document representations in a multi-dimensions space. In this context, we present a novel idea which combines Latent Dirichlet Allocation and Word2vec (providing a consistency metric regarding the partitioned topics) as potential method for limiting the "a priori" number of cluster K usually needed in classical partitioning approaches. We proposed 2 implementations of this idea, respectively able to: (1) finding the best K for LDA in terms of topic consistency; (2) gathering the optimal clusters from different levels of clustering. We also proposed a non-traditional visualization approach based on a multi-agents system which combines both dimension reduction and interactivity.**

***Keywords*-weak signal; clustering topics; word embedding; multi-agent system; vizualisation;**

I. Introduction

For decision-makers, the main objective is to make informed decisions despite the drastic increase in signals transmitted by ever more information systems. Saturation phenomena of the capacities of traditional systems leads to difficulties of interpretation or even to refuse the signals precursor to facts or events. Decision-making is constrained by temporal necessities and thus requires rapidly processing a large mass of information. Also, it must address both the credibility of the source and the relevance of the information revealed and thus requires robust algorithms for weak signal detection, extraction, analysis of the information carried by this signal and the openness towards a wider-scale information context.

Our goal is therefore the detection of precursor signals whose contiguous presence in a given space of time and places anticipates the occurrence of an observable fact. This detection is facilitated by the early information provided by a whistleblower in form of documents which expose proven, unitary and targeted facts but also partial and relating to a triggering event.

At a higher level, this study aims at establishing an investigation procedure able to address the following actions. **A1**: Automatic content analysis with minimal *a priori* information. Identification of relevant information, themes categorization along with coherence indicators. **A2**: Aggregation of knowledge and enrichment of the information. **A3**: Visualization by putting information into perspective by creating representations and dynamic dashboards.

More specifically, the contributions followingly described in this paper essentially tackles these actions, and proposes a solution respectively for (1) detecting the weak signals, (2) extracting the information conveyed by them and (3) presenting the informations in an interactive way. Our system automatically extracts, analyses and put informations into dashboards. It builds indicators for recipients who can also visualize the dynamic evolution of information provided by a multi-agent environment system. Actually, rather than using PCA or tSNE [1] for visualizing our documents in reduced 2D space, we opted for an "attraction/repulsion" multi-agent system into which distances between agents (i.e. documents) are driven by there similarities (regarding extracted features). This approach has the advantage of offering both capabilities of real time evolution and rich interaction to the end user (e.g. by forcing the position of some agents).

The article is organized as follows: first, a review on "weak signals" is given in order to enlighten the context of the study and to underline the multiple definitions found in the literature. Then, a more technical state-of-the-art is provided for topic modeling and word embedding methods which are both involved in our proposed solution. The Section III presents one of our contributions: an "*LDA*[1] augmented with *Word2Vec*" solution for weak signal detection. Some results showing the interest of this approach. Finally, an interactive vizualisation solution is presented which allows revelant management of documents carrying weak signals.

A. A review on "weak signals"

In this constant data growth context, the detection of weak signals has become an important tool for decision makers. Weak signals are the precursors of future events. Ansoff [2] proposed the concept of weak signal in a strategic planning objective through environmental analysis. Typical examples of weak signals are associated with technological developments, demographic change, new actors, environmental change, etc. [3]. Coffman [4] proposed a more specific definition of Ansoff's weak signal as a source that affects a business and its

[1]Latent Dirichlet Allocation

environment. It is unexpected by the potential receiver as much as it is difficult to define due to other signals and noise.

The growing expansion of web content shows that automated environmental scanning techniques (could ou often) outperform research manualy made by human experts [5]. Other techniques combining both automatic and manual approaches have also been implemented [6], [7].

To overcome these limitations, Yoon [7] proposes an automated approach based on a keyword emergence map whose purpose is to define the visibility of words (TF: term frequency) and a keyword emission map that shows the degree of diffusion (DF: document frequency). For the detection of weak signals, some work uses Hiltunen's three-dimensional model [8] where a keyword that has a low visibility and a low diffusion level is considered as a weak signal. On the contrary, a keyword with strong TF and DF degrees is classified as a strong signal.

Kim and Lee [6] proposed a joint approach based on word categorization and the creation of word clusters. It positioned itself on the notions of rarity and anomaly (outliers) of the weak signal, and whose associated paradigm is not linked to existing paradigms. Thorleuchter [9] completed this definition by the fact that the weak signal keywords are semantically related. He therefore added a dependence qualifier.

Like the detection of weak signals, novelty detection is an unsupervised learning task that aims to identify unknown or inconsistent samples of a training data set. Each approach developed in the literature specializes in a particular application such as medical diagnostics [10], monitoring of industrial systems [11] or video processing [12]. In the novelty detection field, references are presented in [13], [14]. In document analysis, the most used and relevant approaches are built by using Latent Dirichlet Allocation (LDA). In this paper, we choose to compare our approach with LDA.

We therefore retain for our study several qualifiers for weak signals. We propose the definition below.

Definition. A *weak signal* is characterized by a low number of words per document and in a few documents (rarity, abnormality). It is revealed by a collection of words belonging to a same and single theme (unitary, semantically related), not related to other existing themes (to other paradigms), and appearing in similar contexts (dependence).

This is therefore a difficult problem since the themes carried by the documents are unknown and the collection of words that make up these themes also. In addition to these difficulties of constructing document classes in an unsupervised manner, there is the difficulty of identifying, via the collections of words that reveal it, the theme related to the weak signal. The analysis must therefore simultaneously make it possible to: 1) discover the themes, 2) classify the documents in relation to the themes, 3) detect relevant keywords related to themes, and finally, 4) it's the main purpose of the study, to discover the keywords related to a *weak signal* theme possibly present. Figure 1 illustrates the processing chain. We focus this paper on the multidimensional clustering problem. We present our attracting/repulsing-based multi-agent model.

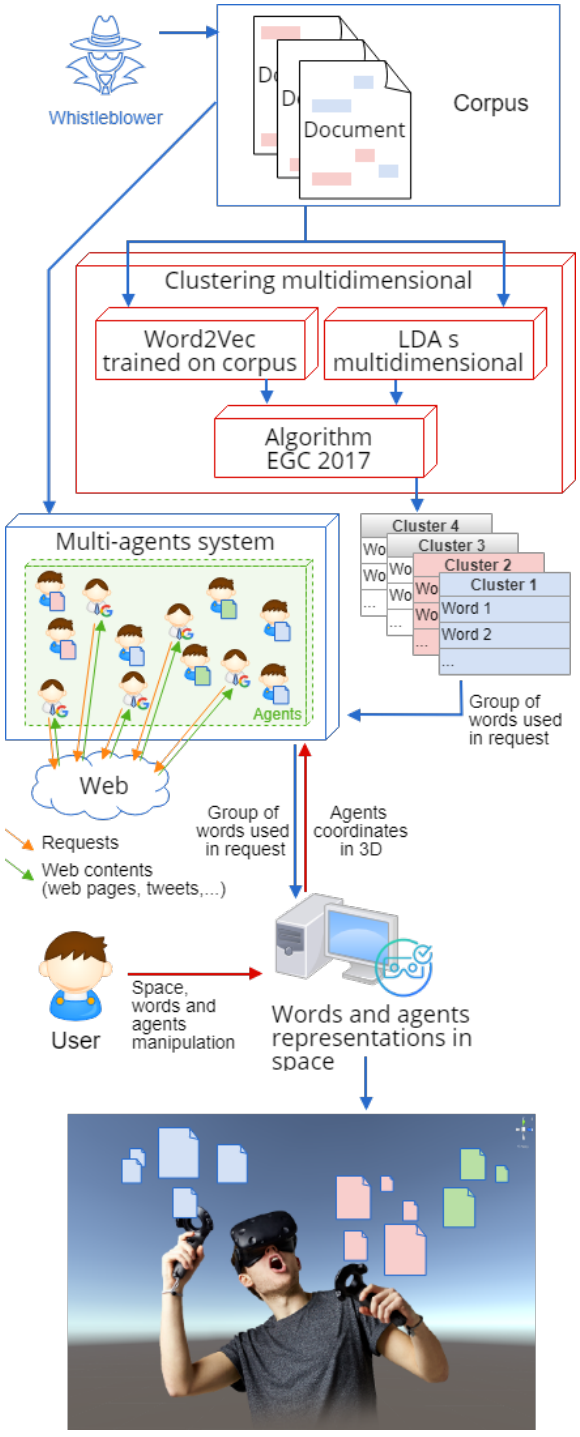

Fig. 1. The system automatically extracts and analyzes the information provided by the whistleblower. The system builds indicators that are put in dashboards for recipients who can also visualize the dynamic evolution of information provided by a multi-agent environment system. This one is used for navigation and document retrieval. Each document is represented by an agent which moves in a 3D environment.

II. State of the art LDA / Word2Vec

This section gives to the readers a state-of-the-art of "topics modeling" and "word embedding" methods which are both involved in our proposed solution. Those 2 methods belong to different, but complementary paradigms in Natural Language Processing.

A. Thematic modeling

Numerous automatic techniques have been developed to visualize, analyze and summarize document collections [15].

In order to manage the data growth explosion, new techniques or tools must be used to process, organize, search, index and browse large collections. Based on machine learning and statistics, topic modeling approaches have been developed to discover word-use patterns that are shared in connected documents [16]. These hierarchical probabilistic models are used to extract underlying topics in documents. They have also been used to analyze contents rather than words such as images, biological data, and survey data [17]. For text analysis and extraction, topic models are based on the bag of words hypothesis.

Different types of "bag of words" exist in the literature. *Latent Semantic Analysis* (*LSA*), *Probabilistic Latent Semantic Analysis* (*PLSA*), *Latent Dirichlet Allocation* (*LDA*) have improved the accuracy of classification in the field of topic discovery and modeling [16], [18].

LDA in the 1990 years was intended to improve the way in which models capture the exchangeability of words and documents compared to previous models *PLSA* and *LSA*: any collection of exchangeable random variables can be represented as a mixture of distributions, often called "infinite" [19].

LDA is an algorithm for the exploration of text based on widely-used Dirichlet process (Bayesian statistics). There are dozens of templates based on *LDA*: extraction of temporal text, author-subject analysis, supervised model of topic, latent Dirichlet co-clustered and bioinformatics relying on *LDA* [20]. In a simplified way, the underlying idea of the process is that each document is modeled as a mixture of topics, and each topic is a discrete probability distribution defining the probability that each word will appear in a given topic. These probabilities on the topics provide a concise representation of the document. Thus *LDA* makes a non-deterministic association between topics and documents [21].

B. Word embedding

Word embedding is the name for a set of language modeling approaches and learning techniques in the field of automatic *Natural Language Processing* (*NLP*) where words are represented by numerical vectors. Conceptually, it is a mathematical integration of a multi-dimensional space where each dimension corresponds to a word in a continuous vector space of much smaller dimension.

Methods to generate this mapping include reduction of dimensionality on the co-occurrence matrix of the words [22], probabilistic models [23], explicit representation according to the context in which the words appear [24] and neural networks [25] like *Word2Vec*.

Word embedding relies on the fact that words are represented as vectors, characteristic of the contextual relationships that connect them through their (neighborhood) context. It is then possible to define the similarity value between two words (later called *w2vSim* in *Word2Vec*). A value close to 1 indicates that the words are very close to each other (i. e. have a similar context) and therefore has a strong semantic link. Conversely, 0 indicates words that are little used in similar contexts.

Word and sentence embedding, when used as the underlying input representation, has significantly increased performance in *NLP* tasks such as syntax parsing [26] and sentiment analysis [27].

III. LDA augmented with Word2Vec

This section describes our contribution which consists in the use of a *Word2Vec* based criterion to filter/select most coherent clusters among those provided by LDA ran at different K levels. To facilitate the understanding of our approach, Figure 2 illustrates the different steps described here below:

- First of all, the upper left part of figure illustrates how we generated augmented corpora carrying weak signals. This generation is performed over original documents extracted from Wikipedia;
- The analysis and the extraction of weak signals are based on a joint *LDA* (I-Topic Modeling) / *Word2Vec* (II-Word Embedding) approach. LDA is therefore applied aver all the documents with different value of K (number of clusters) in order to obtain a set of partitions linked together in the form of a tree structure. This tree is then pruned (III-Pruning) using a coherence criterion to identify a subset of clusters where at least one of them is likely to contain the keywords of the weak signal.
- Finally (IV-Sorting), the cluster carrying the weak signal is identified. The same consistency criterion is used but only on rare words in the cluster.

LDA is a unsupervised classification method and therefore does not allow to associate a label with the found clusters. Moreover, it is difficult to discern the coherence of each cluster. For this it is necessary to define an indicator, this is the subject of the next section. Finally, it is complex to assess whether the topics have actually been found since a word could be justified in several clusters.

A. Intra-cluster consistency metric

The first proposed indicator of local consistency relies on a word embedding method, *Word2Vec* [28]. It proposes to qualify the intrinsec semantic similarity of a cluster of words (topics) in the context of the corpus of documents. This first indicator is defined as follows:

$$indicator\ I_1 = \sum_{w \epsilon E} w2vSim(w_i, w_j) \tag{1}$$

with indicator I_1 being the sum of the similarity values of all combinations of pairwise words in each cluster, with

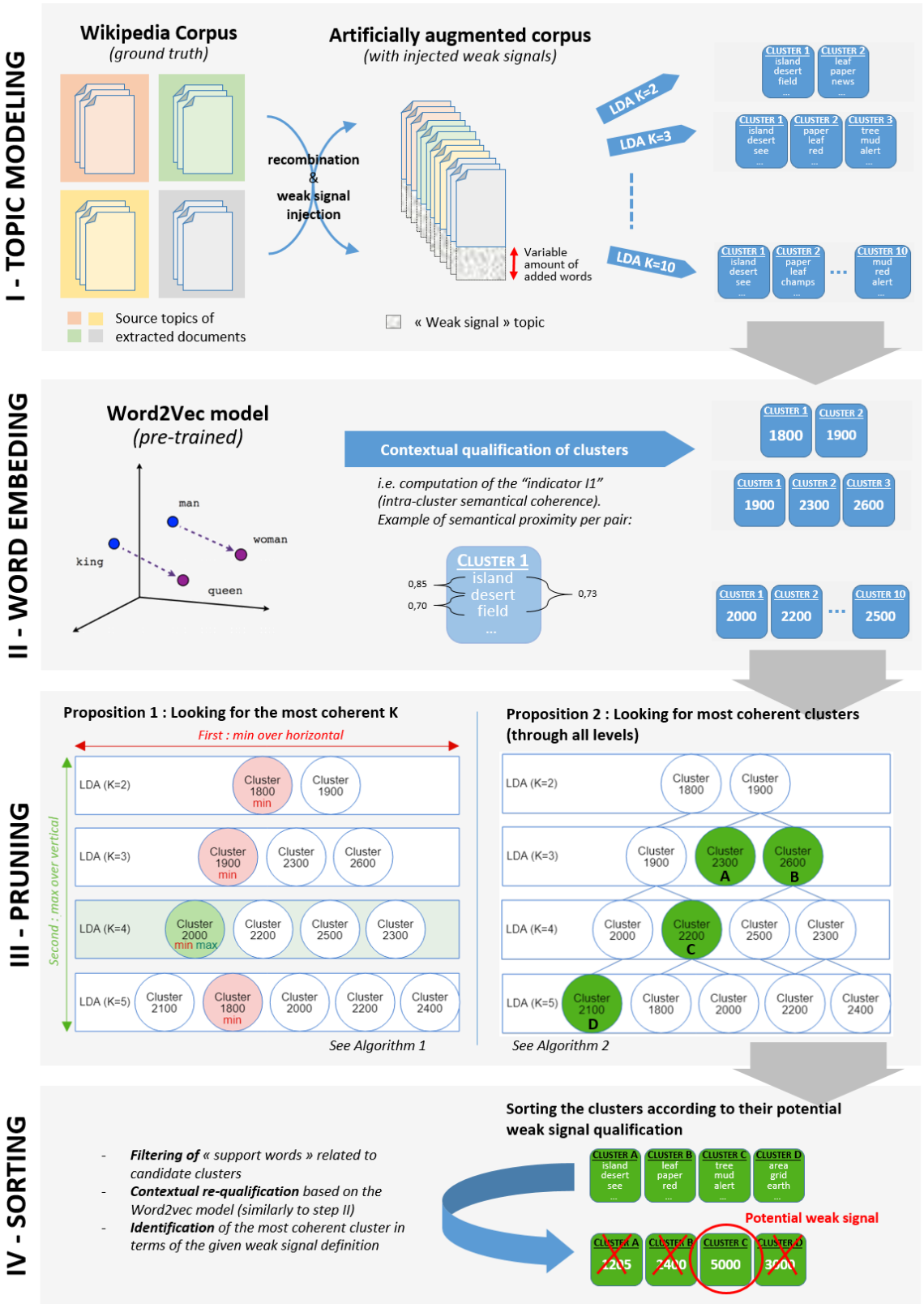

Fig. 2. Overview of our approach to detect a category qualified as weak signal among a corpus of documents belonging to more classical categories.

$E = \{w_1, w_2, \ldots, w_{100}\}$ is the set of the first 100 words supporting the cluster and $w2vSim$ represents the similarity measure (cosine distance defined in *Word2Vec* [28]). The greater the value, the more words the cluster contains that are regularly used together.

This indicator I_1 was proposed to be used over clusters made by *LDA* algorithm with different number of clusters/topics (K) in order to obtain several partitions, which can be represented in the form of a tree structure. It should be noted that *LDA* organizes discovered clusters during the different iterations in random order. An additional step presented in section III-C is therefore necessary to build the tree structure.

In order to evaluate the obtained partitions, two algorithms are proposed based on the previous indicator: 1) a first algorithm (described in Section III-B) aiming at searching the number of clusters (parameter K) leading to a partitioning by *LDA* the most coherent possible; 2) an algorithm (described in Section III-C) which, in a more advanced way, by an in-depth tree analysis, combines the best clusters returned by *LDA* on all partitions (or values of K tested).

B. Best K with the coherence criterion

The Algorithm 1 consists of searching for the level of the tree structure giving the most consistent clusters in the sense of this indicator I_1. On each level, we calculate the minimum value of this indicator for all the clusters present on the level. The K level chosen (and therefore the number of relevant clusters within the meaning of the criterion) corresponds to the one with the highest minimum value (illustrated in Figure 2):

$$\underset{k}{Argmax}(min(LDA(k))) \tag{2}$$

Algorithm 1 Recovery of the ID of the optimal K level

Require: P = List of the number of clusters requested: {2...K}
  $bestK \leftarrow 0$ = ID of K level
  $bestScoreK \leftarrow$ Min(LDA($bestK$))
  for all $k \in P$ **do**
    if Min(LDA(k)) $> bestScoreK$ **then**
      $bestK \leftarrow k$
      $bestScoreK \leftarrow$ Min(LDA(k))
    end if
  end for
  return $bestK$

The result obtained by the $bestK$ variable corresponds to the *LDA* level for which the clusters are most relevant in the sense of indicator I_1.

C. Most relevant clusters through all levels

In order to build the tree structure, it is necessary to evaluate a similarity link between clusters of different levels. This is calculated using a resemblance indicator based on Bhattacharyya distance defined as follows :

$$indicator\ I_2 = \sum_{w \epsilon E} \sqrt{p_{w_i} \cdot q_{w_i}} \tag{3}$$

For the set E defined by the common words w present in both a cluster C_K of level K and a cluster C_{K+1} of level $K+1$ (K corresponding to the *LDA* level), we calculate the sum of the probabilities products, p_{w_i} and q_{w_i}, of each word present in the respective clusters C_K and C_{K+1}.

It is then possible to extract on all the tree structure the most relevant clusters within the meaning of the coherence criterion, indicator I_1, and similarity relationships, indicator I_2, between two clusters of K and $K+1$ level. For this purpose, we propose to prune the tree recursively following as ordered exploration based on the indicator I_2. During this process, each newly encountered cluster, retained as relevant, leads to the withdrawal in the tree structure of all its parent and son clusters. Relationships between clusters (described by the indicator I_2) are only considered beyond an arbitrarily defined threshold (illustrated in Figure 2). The algorithm 2 formalizes this heuristic where the $Parents(C_K)$ and $Childs(C_K)$ routines retrieve, respectively, the list of parent and son clusters of the C_K cluster. As results, we obtain a list of relevant clusters that are not connected in the sense of the indicator I_2.

Algorithm 2 Retrieving relevant clusters in the *LDA* tree structure

Require: T = List of clusters in the tree structure *LDA* sorted by consistency value
  $retainedClusters \leftarrow \{\}$ = List of relevant clusters
  while Size(T) > 0 **do**
    $bestCluster \leftarrow Max(T)$
    $retainedClusters \leftarrow retainedClusters + \{bestCluster\}$
    for all $t \in$ Parents($bestCluster$) **do**
      $T \leftarrow T - \{t\}$
    end for
    for all $t \in$ Childs($bestCluster$) **do**
      $T \leftarrow T - \{t\}$
    end for
  end while
  return $retainedClusters$

Algorithm 2 has following properties :

$$\begin{cases} \overline{I_1^i}^{Alg2} \geqslant \overline{I_1^i}^{LDA_K} & \forall K \\ \underset{i}{max}\ C_i^{Alg2} \geqslant \underset{i}{max}\ C_i^{LDA_K} & \forall K \end{cases} \tag{4}$$

The average of the semantic coherences of the found cluters is increased.

In order to evaluate the performance of the approach, it is necessary to confront them experimentally with the use of *LDA* alone. This assessment is discussed in the next section.

IV. Experimentation

For our experiments, we focused on a subset of documents extracted from the French Wikipedia (snapshot of 08/11/2016) over 5 different categories : Economy, History, Informatics, Health and Law. Wikipedia articles are organized in a particular tree structure and the crawling was done by exploring hyperlinks as branches until leaves are reached.

For the need of our experimentation and to allow an evaluation by some metrics, we need to generate a new test

corpus that involves some simulated "weak signal". To this purpose, we extracted statistics from our initial corpus and then described three groups of words, as shown in Figure 3: 1) common words, belonging to 3 or more categories (first 12% of words in the corpus sorted by occurrence); 2) words belonging to two categories; 3) words belonging to a single category.

	HISTORY	ECONOMY	INFORMATICS	HEALTH	LAW
HISTORY	394286	49387	16007	35752	14523
ECONOMY		80868	12664	5204	3669
INFORMATICS			60614	2196	931
HEALTH				74859	1209
LAW					14920

Words encounted in 1 theme only : 625547
Words encounted in 2 themes 141542
Words encounted in 3 theme on more : 110441

Fig. 3. Presentation of the corpus extracted from Wikpedia. Words encountered in 3 themes, also called "common words", represent about 12 percent of the words in the corpus sorted by occurrence.

Figure 6 illustrates more in detail how the test corpus is generated. It involves common and non-common words which are identified by a study of co-occurrence among all the documents of the corpus. Words chosen to model the weak signal are picked-up from "Law" related documents and inserted, after filtering, in variable quantities of documents of the corpus. The word distributions are respected during insertion, and only the stop-words are deleted.

For this test, we used a *Word2Vec* model pre-trained on the French Wikipedia corpus (Dump of 07/11/2016). The inference is made over the sum of the cluster belonging likelihoods, the words falling within the single category defined above. The topic that has the greatest value is chosen. We indicate in the table if the *weak signal* topic is detected by algorithm 2 in all the selected clusters, and if finally the cluster carrying the weak signal is detected.

All data (documents from French Wikipedia and pre-trained *Word2Vec* model) used in this work are publicly available following these links: *https://zenodo.org/record/3260046 [29]*, *https://zenodo.org/record/162792 [30]*

Each dataset consists of 250 documents from each topic. We insert in a variable number of documents 3 groups of 4 words belonging to the Law topic only. The latter acts as a *weak signal*. The threshold for determining the tree structure is set at 0.75. It is chosen empirically after several experimentations (parameter sampling) on a subset of dataset. In our experiments, impacts of this value are not very sensitive when it belongs to the interval [0.6-0.9]. It allows obtaining the best coherence values for clusters as well as the best detection of weak signal clusters. This value affects the number of clusters detected. The words of the *weak signal* are inserted respecting the distributions previously calculated on the corpus of documents. The number of documents with the *weak signal* varies from 100 to 800 in steps of 50 documents. We perform this test 10 times.

The results obtained (see Figure 4), show the effectivness of algorithm 2 event for a very small number of words injected from the Law category (*weak signal*) compared to each LDA for K=2...8. For a detection level of 8 on 10 tests, it is necessary to inject 0.82% of the words of the *weak signal* topic in relation to the total words of the corpus. The words of the *weak signal* are injected into a document in the form of 3 series of 4 words (12 words per document). 0.82% corresponds to 3′600 words (12 words injected into 300 documents). Each time we found the weak signal cluster, we are looking for the one with the highest similarity coherence value. In Figure 5, we show that the algorithm 2 can detect the weak signal cluster with most coherence value through all level of LDA for K=2...8. The *LDA* algorithm alone sometimes gives a partition where the *weak signal* cluster is present (with a lower similarity coherence value). However, it is necessary to identify this cluster later on. This test therefore shows the interest and contribution of this study in the detection of a weak signal by a joint *LDA*/*Word2Vec* approach.

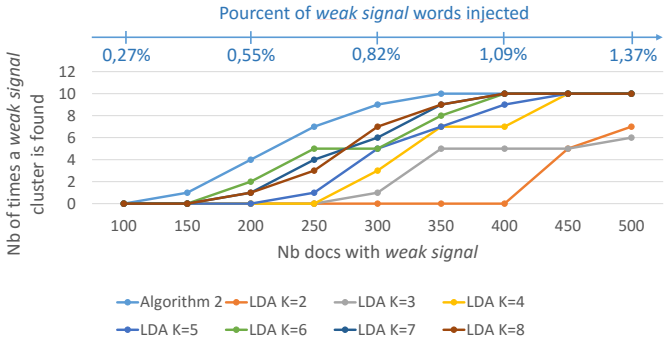

Fig. 4. Result of the algorithm 2 compared to 7 original *LDAs* parameterized with K from 2 to 8 on the detection of a weak signal cluster in the results clusters. For each document, we insert 3 series of 4 words from the Law category (*weak signal*).

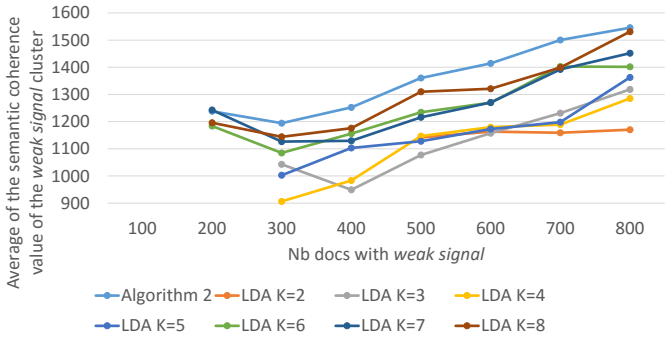

Fig. 5. Result of the algorithm 2 compared to 7 original *LDAs* parameterized with K from 2 to 8 on find the weak signal cluster with the most coherence value through all levels.

V. Visualization

The visualization involves that documents themselves along with the words spotted be part of the "*weak signal*" topic. Those keywords are indeed used to fuel a self-organising Multi-Agent System (MAS), where document agents are animated by attracting/repulsing forces based on semantic similarities. This MAS can be described as follows: 1) new document agents are spawned as a result of queries made over a search engine; 2) the agents are constantly moving, which allows an active spatial reorganization of the documents and so the visible clusters also; and 3) human interactions are possible by manually forcing the position of particular documents agents. Figure 7 shows the concept. The system is actively searching for documents related to the *weak signal* topic, progressively increasing the size of the corpus by spawning new documents and discovering other related words possibly picked-up for the same cluster. The methodological approach is intended to be consistent with the one adopted, for instance, by journalists, who first rely on unitary and targeted facts/documents, then attempt to consolidate them and assess their relevance by exploring other sources. These make it possible to open up to a broader informational context.

Figure 8 shows MAS in action which actively searches for new documents while it is spatially reorganizing the existing document agents into clusters. This model simplifies the problem of mapping a high-dimensional feature space onto a 3D space in order to facilitate the visualization and allows an intuitive user interaction. By forcing the position of agents in space, the agents become automatically some kind of query-agent, letting no choice to the others free agents to rearrange the positions around the fixed one(s).

VI. Conclusion

We present an approach for searching common topics in a corpus of documents and detecting a topic related to a *weak signal* characterized by a small number of words per document and present in few documents. The combination *LDA*/*Word2Vec* as we proposed to implement it, allows us to free ourselves from the arbitrary choice of the K parameter (number of clusters) during partitioning. Two directions were explored: 1) the algorithm 1 aims to find the number of topics leading to a partitioning by *LDA* as consistent as possible; 2) the algorithm 2 which, in a more advanced way, combines the best topics returned by *LDA* on the whole tree structure built when K is varied. This algorithm uses a more relevant indicator to evaluate the similarity link between clusters of different levels of the tree. The goal is to find the set of relevant clusters having the greatest coherence in this tree, whatever the level K. This approach is more suitable for detecting weak signals.

In the context of our study on detecting *weak signal* and issuing alerts, we believe that these *weak signal* deserve to be studied. The information carried by the latter will be correlated with a broader informational context through exploration phases on the networks. The user interacts with the multi-agent system to guide requests on the web.

References

[1] L. van der Maaten and G. Hinton, "Visualizing Data using t-SNE," *Journal of Machine Learning Research*, 2008.

[2] H. I. Ansoff, "Managing Strategic Surprise by Response to Weak Signals," *California Management Review*, vol. XVIII, no. 2, pp. 21–33, 1975.

[3] G. S. Day and P. J. H. Schoemaker, "Scanning the Periphery," *Harvard Business Review*, vol. 83, no. 11, pp. 135–148, 2005.

[4] B. Coffman, "Weak signal research, part I: Introduction," *Journal of Transition Management*, 1997.

[5] R. Decker, R. Wagner, and S. W. Scholz, "An internetâbased approach to environmental scanning in marketing planning," *Marketing Intelligence & Planning*, vol. 23, no. 2, pp. 189–199, 2005.

[6] J. Kim and C. Lee, "Novelty-focused weak signal detection in futuristic data: Assessing the rarity and paradigm unrelatedness of signals," *Technological Forecasting and Social Change*, vol. 120, no. June 2016, pp. 59–76, 2017.

[7] J. Yoon, "Detecting weak signals for long-term business opportunities using text mining of Web news," *Expert Systems with Applications*, vol. 39, no. 16, pp. 12 543–12 550, 2012.

[8] E. Hiltunen, "The future sign and its three dimensions," *Futures*, vol. 40, no. 3, pp. 247–260, 2008.

[9] D. Thorleuchter and D. Van Den Poel, "Weak signal identification with semantic web mining," *Expert Systems with Applications*, vol. 40, no. 12, pp. 4978–4985, 2013. [Online]. Available: http://dx.doi.org/10.1016/j.eswa.2013.03.002

[10] L. Clifton, D. A. Clifton, P. J. Watkinson, and L. Tarassenko, "Identification of Patient Deterioration in Vital-Sign Data using One-Class Support Vector Machines," *Proceedings of the Federated Conference on Computer Science and Information Systems*, no. ii, pp. 125–131, 2011.

[11] A. Ebrahimkhanlou and S. Salamone, "A probabilistic framework for single-sensor acoustic emission source localization in thin metallic plates," *Smart Materials and Structures*, vol. 26, no. 9, 2017.

[12] R. Ramezani, P. Angelov, and X. Zhou, "A fast approach to novelty detection in video streams using recursive density estimation," in *2008 4th International IEEE Conference Intelligent Systems, IS 2008*, vol. 3, 2008, pp. 142–147.

[13] R. Mohammadi-Ghazi, Y. M. Marzouk, and O. Büyüköztürk, "Conditional classifiers and boosted conditional Gaussian mixture model for novelty detection," *Pattern Recognition*, vol. 81, pp. 601–614, 2018. [Online]. Available: https://doi.org/10.1016/j.patcog.2018.03.022

[14] R. Domingues, M. Filippone, P. Michiardi, and J. Zouaoui, "A comparative evaluation of outlier detection algorithms: Experiments and analyses," *Pattern Recognition*, vol. 74, pp. 406–421, 2018. [Online]. Available: https://doi.org/10.1016/j.patcog.2017.09.037

[15] R. M. Nallapati, A. Ahmed, E. P. Xing, and W. W. Cohen, "Joint latent topic models for text and citations," *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2008.

[16] R. Alghamdi and K. Alfalqi, "A Survey of Topic Modeling in Text Mining," *IJACSA) International Journal of Advanced Computer Science and Applications*, vol. 6, no. 1, pp. 147–153, 2015.

[17] D. M. Blei and J. D. Lafferty, "Dynamic topic models," in *Proceedings of the 23rd international conference on Machine learning - ICML '06*, 2006.

[18] T. Hofmann, "Unsupervised learning by probabilistic Latent Semantic Analysis," *Machine Learning*, vol. 42, no. 1-2, pp. 177–196, 2001.

[19] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *Journal of machine learning research : JMLR*, vol. 3, pp. 993–1022, 2003.

[20] Z.-Y. Shen, S. Zhi-Yong, J. Sun, S. Jun, S. Yi-Dong, and Y.-D. Shen, "Collective Latent Dirichlet Allocation," in *Eighth IEEE International Conference on Data Mining*, 2008.

[21] L. Rigouste, O. Cappé, and F. Yvon, "Quelques observations sur le modèle LDA," *Journées internationales d'Analyse statistique des Données Textuelles*, vol. 8, 2006.

[22] O. Levy and Y. Goldberg, "Neural Word Embedding as Implicit Matrix Factorization," *Advances in Neural Information Processing Systems (NIPS)*, 2014.

[23] A. Globerson, G. Chechik, F. Pereira, and P. N. Tishby, "Euclidean Embedding of Co-occurrence Data," *Journal of Machine Learning Research*, vol. 8, pp. 2265–2295, 2007.

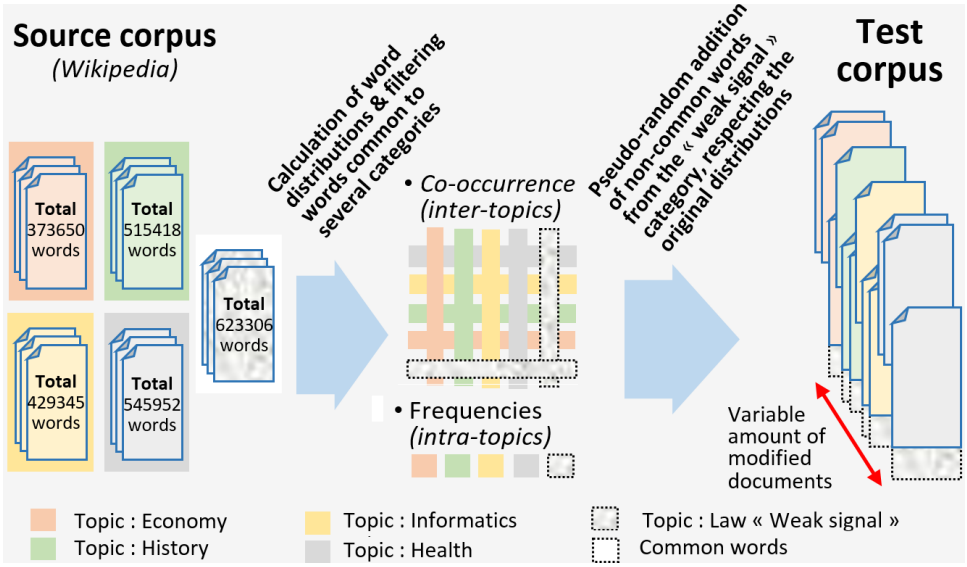

Fig. 6. Test corpus generated by injecting "non-common" words from the "weak signal" category.

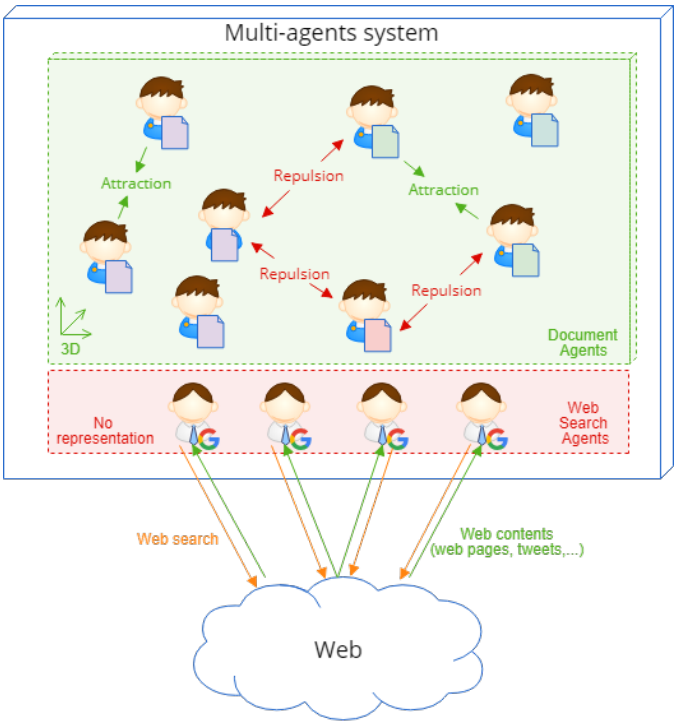

Fig. 7. Representation of multi-agent system. Document agents interact with each others in a attracting/repulsing-based multi-agent system. Search agents request web search engines for find new documents/words related to informations extract in "*weak signal*" documents.

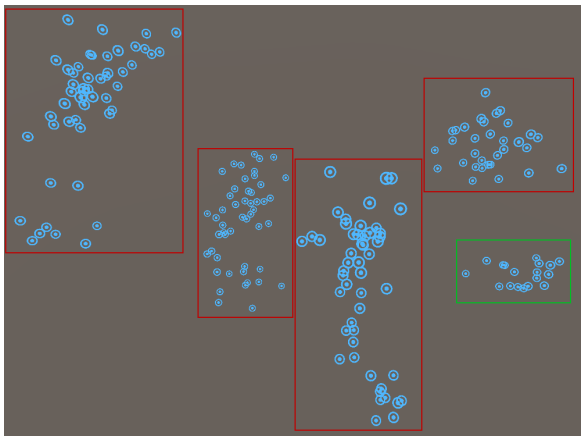

Fig. 8. Our multi-agent system in action showing its ability to self-organize documents in a 3D space but also to make clusters emerge. The red box represents some main topics and green box represents the "*weak signal*" topic.

[24] O. Levy and Y. Goldberg, "Linguistic Regularities in Sparse and Explicit Word Representations," in *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, 2014.

[25] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed Representations of Words and Phrases and their Compositionality," *CrossRef Listing of Deleted DOIs*, pp. 1–9, 2013.

[26] R. Socher, J. Bauer, C. D. Manning, and A. Y. Ng, "Parsing with compositional vector grammars," *ACL 2013 - 51st Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, 2013.

[27] R. Socher, A. Perelygin, and J. Wu, "Recursive deep models for semantic compositionality over a sentiment treebank," *Proceedings of the 2013 conference on empirical methods in natural language processing*, pp. 1631–1642, 2013.

[28] T. Mikolov, G. Corrado, K. Chen, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," *Proceedings of the International Conference on Learning Representations (ICLR 2013)*, pp. 1–12, 2013.

[29] J. Maitre, "A Wikipedia dataset of 5 categories," jun 2019. [Online]. Available: https://zenodo.org/record/3260046

[30] C. Schöch, "A word2vec model file built from the French Wikipedia XML Dump using gensim." oct 2016. [Online]. Available: https://zenodo.org/record/162792