# Recherche de signaux faibles dans un contexte d'investigation numérique

**Julien Maitre* — Michel Ménard* — Alain Bouju* — Guillaume Chiron***

* *L3i*
*La Rochelle Université*
*Avenue Michel Crépeau 17042 La Rochelle, France*

julien.maitre@uni-lr.fr, michel.menard@univ-lr.fr, alain.bouju@univ-lr.fr
guillaume.chiron@univ-lr.fr

RÉSUMÉ. *L'étude présentée s'inscrit dans le cadre du développement d'une plateforme d'analyse automatique de documents associée à un service sécurisé lanceurs d'alerte, de type GlobalLeaks, focalisé sur la révélation de faits/événements/actions en lien avec des problématiques environnementales. Le présent article aborde le problème particulier du regroupement de sujets à plusieurs niveaux à partir de documents multiples, puis de l'extraction de descripteurs significatifs, tels que des listes pondérées de mots. Dans ce contexte, nous présentons une nouvelle idée combinant LDA (en charge du clustering) et Word2vec (fournissant une métrique de cohérence concernant les sujets partitionnés) comme méthode potentielle pour limiter le nombre "a priori" de cluster K habituellement nécessaire dans les approches classiques du partitionnement. Nous avons proposé 2 mises en œuvre de cette idée, respectivement en mesure de : (1) trouver le K optimal pour LDA ; (2) rassembler les clusters optimaux de différents niveaux de clustering.*

ABSTRACT. *This paper is related to a wide project aiming at discoving from different streams of information (i.e. daily publication from the Internet), weak signals possibly sent by whistleblowers. The current 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. In this context, we present a novel idea combining LDA (in charge clustering) 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 optimal K for LDA; (2) gathering the optimal clusters from different levels of clustering.*

MOTS-CLÉS : *Modèle de thèmes, Plongement de mots, LDA, Word2Vec, regroupement*

KEYWORDS: *Topic Modeling, Word Embedding, LDA, Word2Vec, clustering*

## 1. Introduction

Pour les décideurs, l'objectif principal est de prendre des décisions éclairées face à l'augmentation drastique des signaux transmis par des systèmes d'information toujours plus nombreux. Des phénomènes de saturation des capacités de nos systèmes conduisent à des difficultés d'interprétation ou même à refuser les signaux précurseurs de faits ou d'événements. La prise de décision est limitée par les nécessités temporelles et nécessite donc un traitement rapide de la masse d'informations. Etre capable de détecter en un temps fixe, les bons signaux porteurs d'informations utiles dans un contexte de stratégie d'anticipation, est un défi devenu permanent pour de nombreux acteurs économiques. Il est donc nécessaire de développer, sous la forme de plateformes d'investigation (cf. Figure 1), de nouveaux services d'aide à la décision pour les politiques et les organisations chargées de ces activités. Les prises de décision, qui doivent porter à la fois sur la crédibilité de la source d'information et sur la pertinence des informations révélées dans un événement, nécessitent des

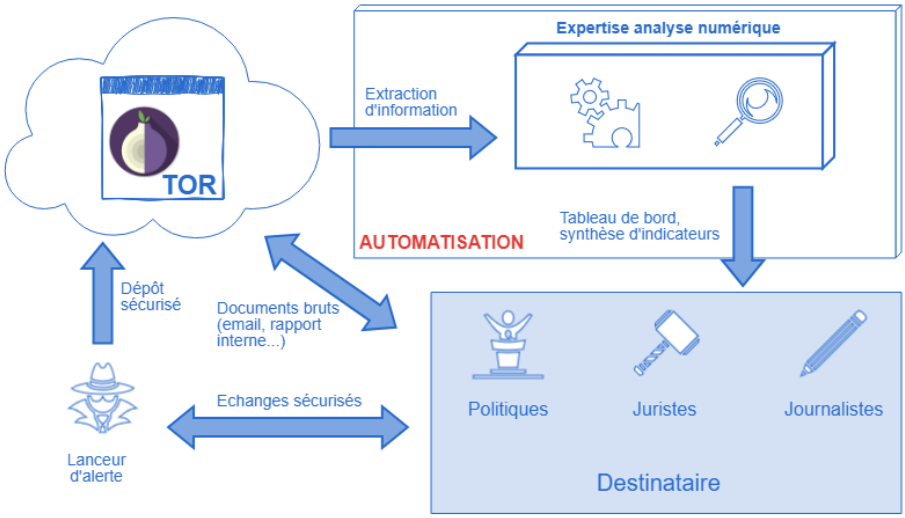



**Figure 1.** *Aperçu de notre plateforme d'investigation. Elle doit répondre au besoin réel des journalistes/politiciens/juristes de disposer d'outils d'investigation (extraction, vérification, corrélation) et de représentation de l'information (synthèse, aide à la décision). Son but est donc de faciliter les expertises indépendantes, de protéger les lanceurs d'alerte et d'aider à détecter les signaux faibles. Les lanceurs d'alerte déposent les premiers documents précurseurs des signaux faibles sur des plateformes numériques construites sur les technologies GlobalLeaks et Tor2Web (e.g. Source Sûre et EULeak). La visualisation et l'interaction avec le système et les différents acteurs (lanceurs d'alerte, journalistes, politiciens, juristes, ...) pourra s'effectuer à l'aide d'un matériel informatique dédié et sécurisé.*

algorithmes robustes pour détecter les *signaux faibles*, extraire, analyser les informations fournies par ce dernier et s'ouvrir à un contexte d'information plus large.

Nous proposons de nous concentrer sur les deux points essentiels suivants: (1) la détection des *signaux faibles* et (2) l'extraction des informations qu'ils véhiculent. La Figure 1 donne notre vision d'une plateforme d'investigation où les acteurs interagissent entre eux par le biais d'outils et de processus sécurisés. Notre but est donc la détection de signaux précurseurs dont la présence contiguë dans un espace de temps et de lieux donné anticipe l'avènement d'un fait observable. Cette détection est facilitée par les premières informations fournies par un lanceur d'alerte sous forme de documents. Ils exposent des faits prouvés, unitaires et ciblés mais aussi partiels et relatifs à un événement déclencheur. Le lanceur d'alerte fournit des informations qui ne sont pas encore détectables / apparentes sur les réseaux sociaux spécialisés. Elles permettent de dessiner le contour des signaux à venir sur les réseaux, facilitant ainsi leur détection et l'extraction des informations qu'ils véhiculent.

La procédure d'investigation proposée repose donc sur la détection de *signaux faibles* à partir des documents reçus.

Trois actions sont donc entreprises :

– Action 1: Analyse automatique de contenus avec un minimum d'*a priori*. Identification des informations pertinentes. Indicateur de cohérence des thèmes obtenus. Détection des *signaux faibles*;

– Action 2: Agrégation de connaissances. Enrichissement de l'information

– Action 3: Visualisation analytique. Mise en perspective de l'information par la création de représentations visuelles et de tableaux de bord dynamique.

L'article s'inscrit principalement dans la première Action. Un système multi-agents valorise l'information de manière visuelle, et grâce à des requêtes, enrichit l'information initiale. Pour compléter cette introduction, nous proposons :

– un premier état de l'art sur cette problématique afin de souligner les définitions plurielles sur lesquelles s'appuient les articles de la littérature pour qualifier un *signal faible*. Nous tentons d'en ressortir une définition commune, et proposons une approche conjointe (globale et contextuelle) dont le but est de s'appuyer sur cette définition pour mettre en évidence le *signal faible*;

– un second état de l'art décrivant les approches de type modèle thématique et plongement de mots afin de déterminer celles qui seront les plus adaptées à la détection des *signaux faibles*.

### 1.1. *Détection des signaux faibles. Etat de l'art et définitions proposées*

Dans ce contexte d'explosion massive de l'information, la détection des *signaux faibles* est devenue un outil important pour les décideurs. Les *signaux faibles* sont les précurseurs des événements futurs. Ansoff (Ansoff, 1975) a proposé le concept de *signal faible* dans un objectif de planification stratégique à travers l'analyse

environnementale. Cet outil est une alternative à la planification stratégique pour les entreprises dans les années 1970 et 1980. Des techniques telles que la prévision et la planification de scénarios, des méthodes de planification stratégique représentatives ont eu tendance à être moins efficaces dans un contexte d'accélération du progrès social où de nombreux facteurs incertains sont apparus. Des exemples typiques de *signaux faibles* sont associés aux développements technologiques, aux changements démographiques, aux nouveaux acteurs, aux changements environnementaux, etc. (Day *et al.*, 2005).

Coffman (Coffman, 1997) a proposé une définition plus précise du *signal faible* d'Ansoff. Il le définit comme une source qui affecte l'environnement des entreprises et ses activités. Il est inattendu pour le récepteur potentiel autant qu'il est difficile à définir en raison des autres signaux et du bruit. Le *signal faible* est une occasion pour les entreprises d'apprendre, de se développer et d'évoluer dans leur environnement.

D'autres travaux théoriques sur les concepts de *signaux faibles* ont été menés par Hiltunen (Hiltunen, 2008). Ce dernier définit le *signal faible* dans un espace tridimensionnel : (1) "signal" qui correspond à la visibilité du signal, (2) "problème" qui représente le nombre d'événements liés au signal et (3) "interprétation" qui est le facteur de compréhension du futur signal par son récepteur.

L'expansion croissante du contenu Web montre que les techniques automatisées d'analyse de l'environnement surpassent la recherche par l'intermédiaire d'experts humains (Decker *et al.*, 2005). D'autres techniques combinant les approches automatiques et manuelles ont également été mises en œuvre (Kim *et al.*, 2017; Yoon, 2012).

Pour surmonter ces limites, Yoon (Yoon, 2012) a proposé une approche s'appuyant sur une carte d'émergence de mots-clés dont le but est de définir la visibilité des mots (TF : term frequency) et une carte d'émission des mots-clés qui montre le degré de diffusion (DF : document frequency). Pour la détection de *signaux faibles*, certains travaux utilisent le modèle tridimensionnel de Hiltunen (Hiltunen, 2008) où un mot-clé qui a une faible visibilité et un faible niveau de diffusion est considéré comme un *signal faible*. Au contraire, un mot-clé avec de forts degrés TF et DF est classé comme un signal fort. Park (Park *et al.*, 2017) a utilisé cette approche en sélectionnant "smart grid" comme mot-clé de recherche et une période de 3 ans pour aider les décideurs politiques et les parties prenantes à mieux comprendre les questions de réseaux intelligents liées aux technologies, aux marchés et à la conduite de projets plus efficaces.

Kim et Lee (Kim *et al.*, 2017) ont proposé une approche conjointe reposant sur la catégorisation de mots et sur la création de clusters de mots. Elle s'appuie sur des matrices "document/mots-clés", cible un domaine et des sources particulières, et nécessite l'intervention d'experts dans une phase de filtrage préalable des mots-clés. Elle se positionne sur les notions de rareté et d'anomalie (outliers) du *signal faible*, et dont le paradigme associé ne soit pas relié à des paradigmes existants. Il est à ce titre nouveau. Thorleuchter (Thorleuchter *et al.*, 2013) complètent cette définition par le

fait que les mots-clés du *signal faible* sont sémantiquement reliés. Il ajoute donc le qualificatif de dépendance.

Nous retenons donc pour notre étude les qualificatifs suivant des *signaux faibles* : rareté, unitaire, non relié à des paradigmes existants, nouveauté, anormalité, mots-clés sémantiquement reliés. Nous proposons les définitions ci-dessous.

**Signal faible**. Un *signal faible* est caractérisé par un faible nombre de mots par document et présents dans peu de documents (rareté, anormalité). Il est révélé par une collection de mots appartenant à un seul et même thème (unitaire, sémantiquement reliés), non relié à d'autres thèmes existants (à d'autres paradigmes), et apparaissant dans des contextes similaires (dépendance).

**Classe ou catégorie, Thèmes et Clusters**. On considère que chaque document de classe inconnue est représenté par un mélange fini de thèmes. Chaque thème est un mélange de mots représentatifs de ce thème qu'il est possible de capturer dans un cluster. Classe et catégorie sont utilisées de manière interchangeable.

**Information présente en entrée du système**. Les données d'entrée du système correspondent à l'ensemble des documents reçus (i.e. transmis par le lanceur d'alerte).

**Information en sortie du système**. En sortie, l'algorithme calcule un ensemble de clusters. Chacun d'eux contient une collection de mots. A chaque cluster sera associé un thème. L'algorithme devra découvrir les mots-clés relevant d'un thème "*signal faible"* éventuellement présent.

Il s'agit donc d'un problème difficile puisque les thèmes portés par les documents sont inconnus et la collection de mots qui composent ces thèmes également. A ces difficultés de construire de manière non-supervisée des classes de documents, s'ajoute celui d'identifier, via la collection de mots qui le révèle, le thème relatif au *signal faible*. L'analyse des documents reçus doit donc simultanément permettre de :

– découvrir les thèmes,

– classer les documents relativement aux thèmes,

– détecter les mots-clés pertinents relatifs aux thèmes,

– et enfin, c'est la finalité même de l'étude, de découvrir les mots-clés relevant d'un thème "*signal faible*" éventuellement présent.

La Figure 2 illustre la chaîne de traitement. Nous nous focalisons dans cet article sur le problème du clustering multi-niveaux. Nous présentons notre modèle multi-agents construit sur un schéma d'attraction/répulsion.

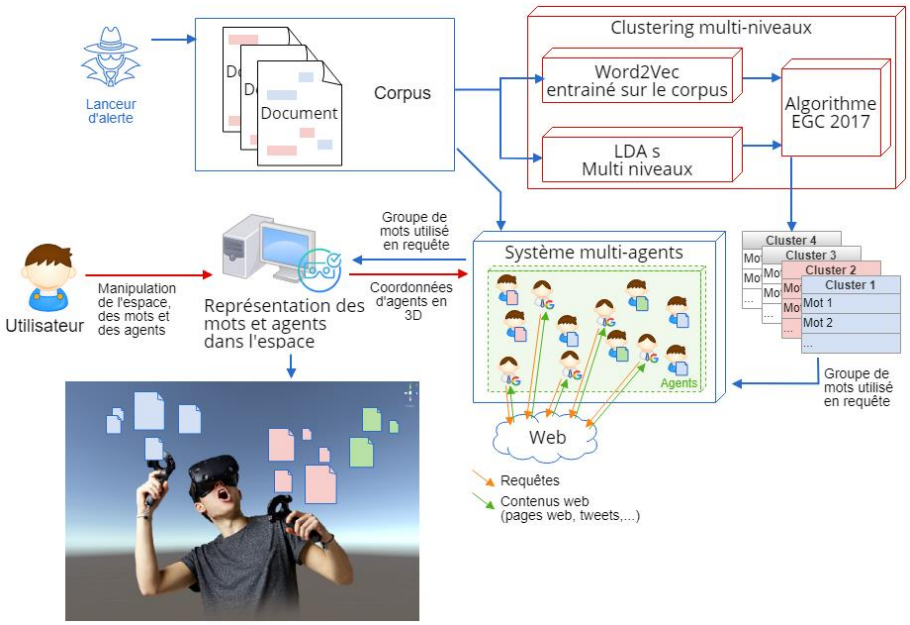



**Figure 2.** Le système extrait et analyse automatiquement les informations fournies par le lanceur d'alerte. Il construit des indicateurs placés ensuite dans des tableaux de bord pour analyse et consultation. Ceux-ci sont présentés de manière dynamique grâce à un système multi-agents dans un objectif de navigation et de recherche documentaire. Dans ce but, chaque document est représenté par un agent se déplaçant dans un environnement 3D.

## 2. Etat de l'art LDA / Word2Vec

### 2.1. *Approches de type modèle thématique*

De nombreuses techniques automatiques ont été mises au point pour visualiser, analyser et résumer des collections de documents (Nallapati *et al.*, 2008).

En s'appuyant sur l'apprentissage automatique et les statistiques, des approches de type "modèle thématique" ont été développées pour découvrir les modes d'utilisation des mots qui sont partagés dans des documents connectés (Alghamdi *et al.*, 2015). Ces modèles sont utilisés pour extraire les thèmes sous-jacents des documents. Ils ont également été adoptés pour analyser le contenu plutôt que des mots tels que des

images, des données biologiques et des données d'enquêtes (Blei *et al.*, 2006). Pour l'analyse et l'extraction de texte, les modèles thématiques se fondent sur l'hypothèse de sac de mots (i.e. les informations sur l'ordre des mots sont ignorées).

Plusieurs approches de type "sac de mots" existent dans la littérature. Citons *l'analyse sémantique latente* (*LSA*), *l'analyse sémantique latente probabiliste* (*PLSA*), *l'allocation de Dirichlet latente* (*LDA*) qui ont amélioré la précision de la classification dans le cadre de la découverte et de la modélisation de thèmes (Hofmann, 2001; Alghamdi *et al.*, 2015).

Au cours des années de 1990, *LDA* a eu pour but d'améliorer la façon dont les modèles saisissent l'interchangeabilité des mots et des documents par rapport aux précédents modèles *PLSA* et *LSA*: toute collection de variables aléatoires échangeables peut être représentée comme un mélange de distributions, souvent un mélange "infini" (Blei *et al.*, 2003).

*LDA* est un algorithme d'exploration de texte s'appuyant sur un processus de Dirichlet (statistiques bayésiennes). Il existe plusieurs modèles construits sur *LDA*: extraction de texte temporel, analyse sujet-auteur, modèle supervisé de thèmes et Dirichlet latent co-clusterisé reposant sur *LDA* (Shen *et al.*, 2008). De manière simplifiée, l'idée sous-jacente du processus est que chaque document est modélisé comme un mélange de thèmes, et que chaque thème est une distribution de probabilité discrète définissant la probabilité que chaque mot apparaisse dans un thème donné. Ces probabilités sur les thèmes fournissent une représentation concise du document. Ainsi, *LDA* établit une association non-déterministe entre les thèmes et les documents (Rigouste *et al.*, 2006).

### 2.2. *Word embedding*

Le plongement de mots (*word embedding*) est le nom donné à un ensemble d'approches de modélisation linguistique et de techniques d'apprentissage dans le domaine du *traitement automatique du langage naturel* (NLP) où les mots sont représentés par des vecteurs de nombres. Conceptuellement, c'est une intégration mathématique d'un espace multidimensionnel, où chaque dimension correspond à un mot, dans un espace vectoriel continu de dimension beaucoup plus faible.

Les méthodes pour générer cette cartographie incluent les réseaux de neurones (Mikolov, Sutskever, *et al.*, 2013), la réduction de la dimensionnalité sur la matrice de co-occurrence des mots (Levy *et al.*, 2014a), les modèles probabilistes (Globerson *et al.*, 2007) et la représentation explicite en fonction du contexte dans lequel apparaissent les mots (Levy *et al.*, 2014b).

Le plongement de mots repose sur le fait que les mots sont représentés comme des vecteurs, caractéristiques des relations contextuelles qui les relient entre eux par l'intermédiaire de leur contexte (de voisinage). Il est alors possible de définir la valeur de similarité entre deux mots (appelée plus tard *w2vSim*). Une valeur proche de 1 indique que les mots sont très proches les uns des autres (c'est-à-dire qu'ils ont un

contexte similaire) et ont donc un lien sémantique fort. Inversement, 0 indique des mots qui sont peu utilisés dans des contextes similaires.

L'intégration de mots et de phrases, lorsqu'elle est utilisée comme représentation d'entrée sous-jacente, a augmenté de manière significative les performances dans les tâches de *NLP* telles que l'analyse syntaxique (Socher, Bauer, *et al.*, 2013) et l'analyse des sensations (Socher, Perelygin, *et al.*, 2013).

## 3. LDA augmenté avec Word2Vec

Cette section décrit notre contribution qui consiste en l'utilisation d'un critère basé sur Word2vec pour filtrer/sélectionner les clusters les plus cohérents parmi ceux fournis par LDA et exécutés à différents niveaux K. Afin de faciliter la compréhension de l'approche, la Figure 3 illustre les différentes étapes.

– Tout d'abord, la partie en haut à gauche de la figure illustre schématiquement la manière avec laquelle nous générons des corpus porteurs de *signaux faibles*, une démarche nécessaire à la validation de notre chaîne de traitement en l'absence de vérité terrain explicite. La génération de corpus est déclinée dans la suite de l'article au travers des différents tests sur des données artificielles et sur des données proches du réel;

– Pour l'analyse des documents et l'extraction du *signal faible*, nous adoptons une approche conjointe LDA/Word2vec (illustrée sur les lignes I-Topic Modeling et II-Word embedding de la figure) que nous justifions. Nous appliquons LDA sur l'ensemble des documents, tout en faisant varier le nombre de clusters, afin d'obtenir un ensemble de partitions reliées entre-elles sous la forme d'une arborescence. Celle-ci est élaguée (ligne III) grâce à un critère de cohérence afin de dégager un sous-ensemble de clusters où au moins l'un d'entre-eux est susceptible de contenir les mots-clés du *signal faible*.

– Enfin (ligne IV), le cluster porteur du *signal faible* est identifié. Le même critère de cohérence est utilisé mais uniquement sur les mots rares du cluster.

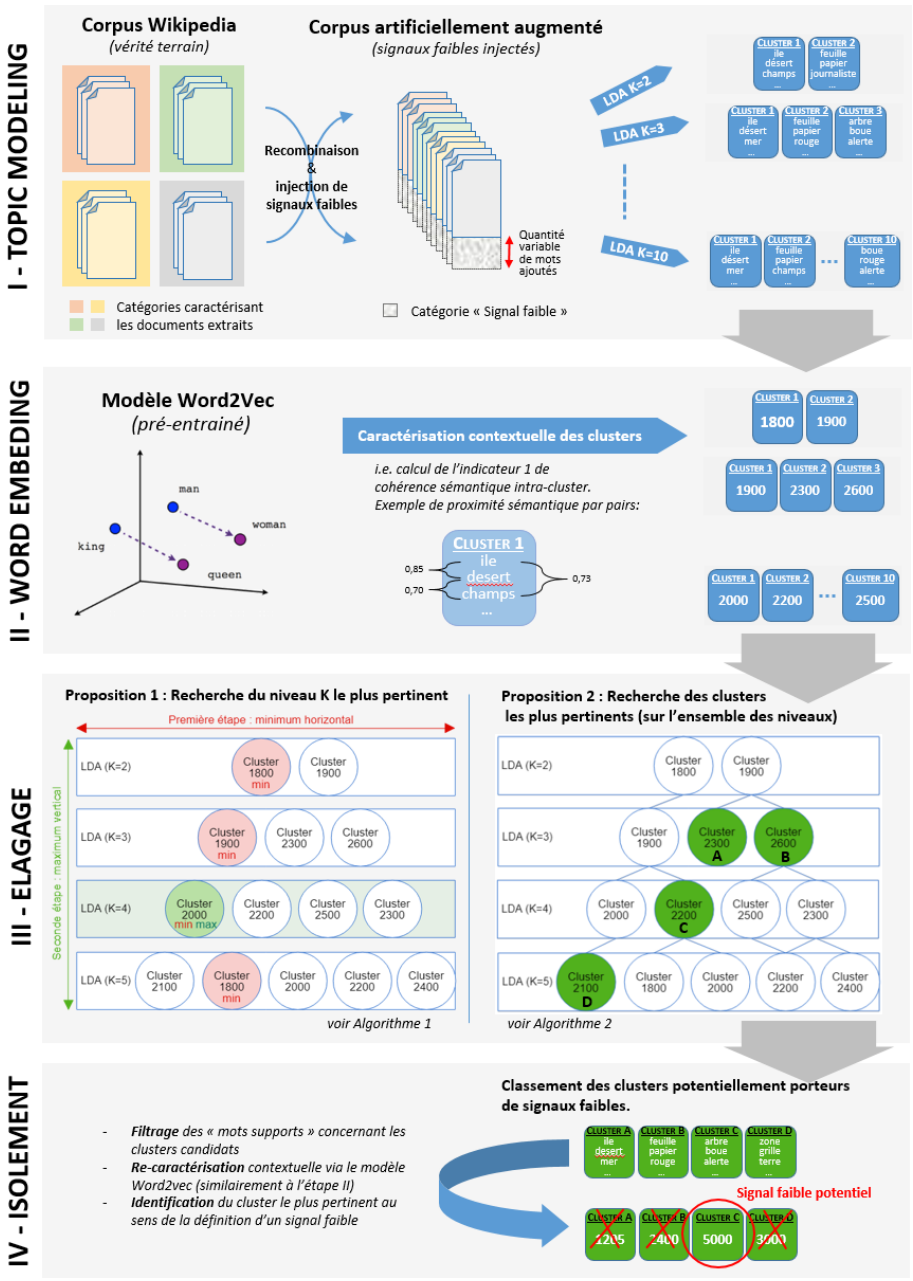



**Figure 3** Les différentes étapes pour extraire les *signaux faibles*

## 4. Expérimentation

Pour illustrer l'utilisation de LDA, nous nous sommes concentrés sur un sous-ensemble de documents extrait de la version française de Wikipedia (snapshot du 08/11/2016) sur 5 catégories différentes: Economie, Histoire, Informatique, Médecine et Droit. Les articles de Wikipédia sont organisés dans une arborescence particulière et le parcours s'est effectué en explorant des hyperliens sous forme de branches jusqu'à ce que les feuilles soient atteintes.

Pour le besoin de notre expérimentation et pour permettre une évaluation par certaines métriques, nous avons besoin de générer un nouveau corpus de test qui implique un "*signal faible*" simulé. Pour ce faire, nous avons extrait des statistiques de notre corpus initial puis décrit trois groupes de mots, comme le montre la Figure 4 : 1) mots communs, appartenant à 3 catégories ou plus (premiers 12% des mots du corpus triés par occurrence) ; 2) mots appartenant à deux catégories ; 3) mots appartenant à une seule catégorie.

| | HISTOIRE | ECONOMIE | INFORMATIQUE | MEDECINE | DROIT |
|---|---|---|---|---|---|
| HISTOIRE | 394286 | 49387 | 16007 | 35752 | 14523 |
| ECONOMIE | | 80868 | 12664 | 5204 | 3669 |
| INFORMATIQUE | | | 60614 | 2196 | 931 |
| MEDECINE | | | | 74859 | 1209 |
| DROIT | | | | | 14920 |

Total des mots rencontrés dans 1 seul thème : 625547

Total des mots rencontrés dans 2 thèmes : 141542

Total des mots rencontrés dans 3 thèmes et + : 110441

**Figure 4** Présentation du corpus extrait de Wikipédia. Les mots rencontrés dans 3 thèmes, aussi appelés "mots communs", représentent environ 12% des mots du corpus triés par occurrence.

La Figure 5 illustre plus en détail comment le corpus de test est généré. Il s'agit de mots communs et non communs qui sont identifiés par une étude de cooccurrence entre tous les documents du corpus. Les mots choisis pour modéliser le *signal faible* sont repris à partir de documents liés au "Droit" et insérés, après filtrage, dans des quantités variables de documents du corpus. Les distributions statistiques de mots du thème "Droit" sont respectées lors de l'insertion et les mots outils sont supprimés.

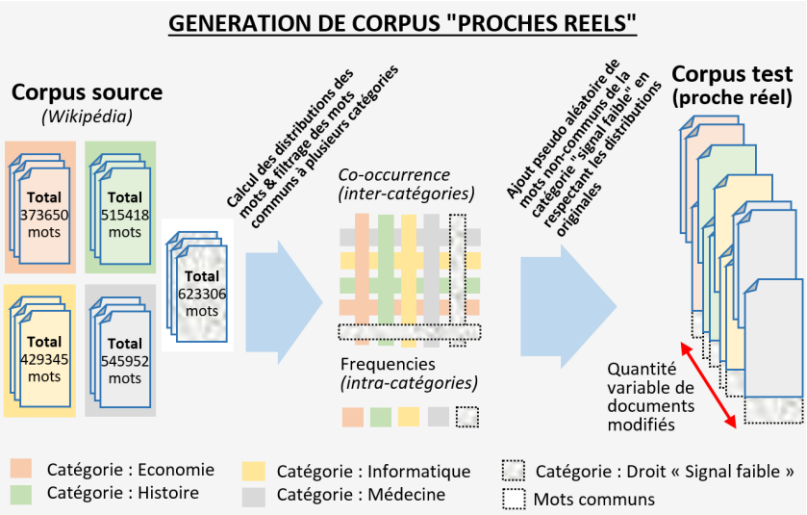



**Figure 5** Méthode de génération du corpus "proche du réel" qui consiste en l'injection de mots dits "non-communs" empruntés au thème "*signal faible"* (Droit en l'occurence) en respectant la distributions originale. Ces mots non-communs sont identifiés par une étude de co-occurrence entre thèmes.

Dans le cadre de ce test, nous utilisons un réseau Word2vec pré-entrainé sur le corpus Wikipédia (Dump du 07/11/2016). L'inférence est effectuée à partir de la somme des degrés d'appartenance des mots du cluster, les mots relevant de la catégorie unique définie précédemment. On retient le thème qui possède la plus grande valeur.

Toutes les données (documents de Wikipédia français et modèle Word2Vec préformé) utilisées dans ce travail sont accessibles au public en suivant ces liens[1] : *https://www.dropbox.com/s/8uilidilpigxm0h/dataCategorizedUTF8.zip*, *https://zenodo.org/record/162792*

Chaque jeu de données est composé de 250 documents de chaque thème. Nous insérons dans un nombre de documents variable 3 groupes de 4 mots appartenant au thème droit uniquement. Ce dernier fait office de "*signal faible*". Le seuil pour la détermination de l'arborescence est fixé à 0,75. Les mots du "*signal faible*" sont insérés en respectant les distributions précédemment calculées sur le corpus de documents. Le nombre de documents portant le "*signal faible*" varie de 50 à 750 par pas de 50 documents. Nous réalisons ce test 10 fois.

1. Afin de ne pas nuire à l'anonymat du présent document, ce lien est temporaire, n'est fourni qu'aux examinateurs et n'est utilisé que pour l'évaluation des contributions du présent document - si le document est accepté, l'ensemble des données sera transféré dans un dépôt de données ouvert et rendu accessible au public.

Les résultats obtenus (voir Figure 6) montrent la robustesse de l'**Erreur ! Source du renvoi introuvable.** même pour un très petit nombre de mots injectés de la catégorie Droit (*signal faible*) comparé à chaque LDA pour K=2...8. Pour un niveau de détection de 8 sur 10 tests, il est nécessaire d'injecter 0,82% de mots du thème *signal faible* par rapport au total des mots du corpus. Les mots du *signal faible* sont injectés dans un document sous la forme de 3 séries de 4 mots, soit 12 mots par document. 0.82% correspond à 3600 mots (10 mots injectés dans 300 documents). Chaque fois que nous trouvons le cluster *signal faible*, nous recherchons celui qui a

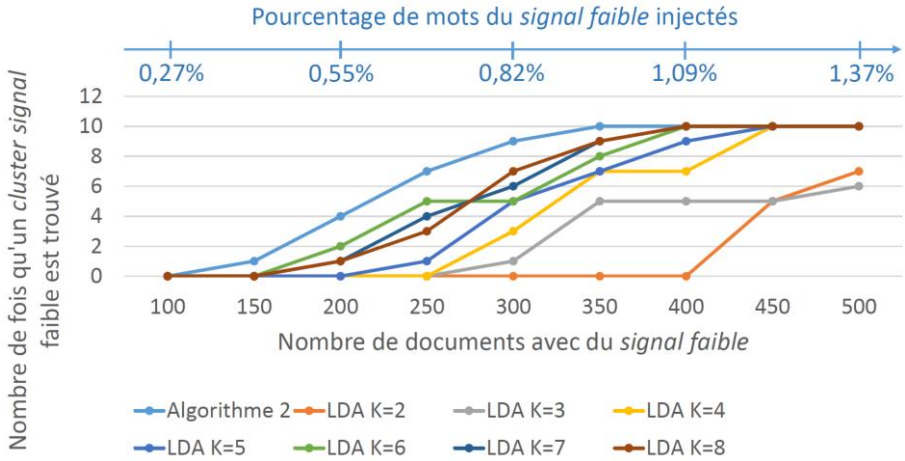



**Figure 6** Résultat de l'algorithme 2 comparé aux 7 LDA originaux paramétrés avec K de 2 à 8 sur la détection d'un cluster *signal faible* dans les clusters de résultats. Pour chaque document, nous insérons 3 séries de 4 mots de la catégorie Droit (*signal faible*).

la valeur de cohérence de similarité la plus élevée. Dans la Figure 7, nous montrons que l'algorithme 2 peut détecter le cluster *signal faible* avec la plus grande valeur de cohérence à travers tous les niveaux de LDA pour K=2. . . 8. L'algorithme LDA seul donne parfois une partition où est présent le cluster *signal faible*. Il est nécessaire cependant d'identifier ce cluster par la suite. Ce test montre donc l'intérêt et la contribution de cette étude quant à la détection d'un *signal faible* par une approche conjointe LDA/Word2vec.

## 5. Visualisation

La visualisation implique que les documents eux-mêmes ainsi que les mots repérés fassent partie du thème *signal faible*. Les mots-clés détectés lors de la phase précédente sont utilisés pour alimenter un système multi-agents auto-organisé (SMA), où les agents document sont animés par des forces d'attraction/répulsion construits

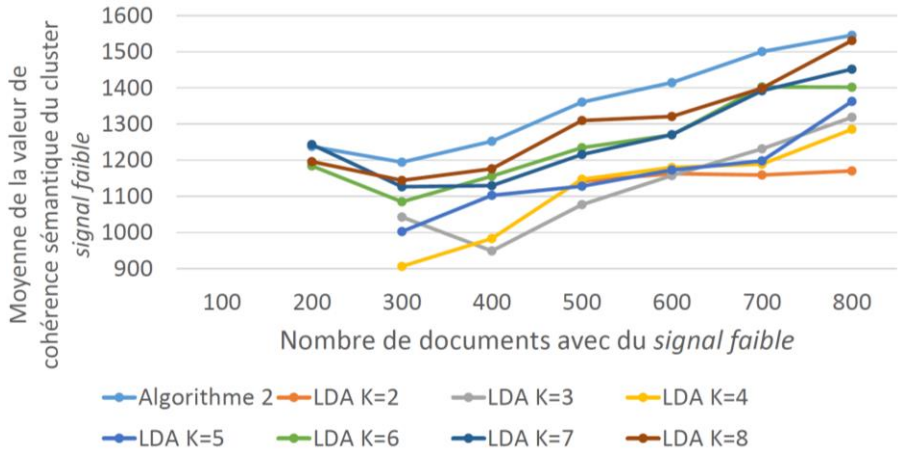



**Figure 7** Résultat de l'algorithme 2 comparé aux 7 LDA originaux paramétrés avec K de 2 à 8 sur trouver le cluster de *signal faible* avec la valeur de cohérence la plus élevée à travers tous les niveaux.

sur des similitudes sémantiques. Ce SMA peut être décrit par les points suivants: 1) de nouveaux agents document sont générés en réponse aux requêtes effectuées sur un moteur de recherche; 2) les agents sont constamment en mouvement, ce qui permet une réorganisation spatiale active des documents et donc aussi des clusters visibles; 3) des interactions humaines sont possibles en forçant manuellement la position d'agents document particuliers. La Figure 8 montre le principe. Le système recherche activement les documents relatifs au thème *signal faible*, augmentant progressivement la taille du corpus par l'apport de nouveaux documents et en découvrant d'autres mots apparentés éventuellement au même cluster. L'approche méthodologique se veut cohérente avec celle adoptée, par exemple, par les journalistes, qui s'appuient d'abord sur des faits et des documents unitaires et ciblés, puis tentent de les consolider et d'évaluer leur pertinence en explorant d'autres sources. Elles permettent de s'ouvrir à un contexte informationnel plus large.

La Figure 9 montre le SMA en action recherchant activement de nouveaux documents tout en réorganisant spatialement les agents document existants en clusters. Ce modèle simplifie le problème du mappage d'un espace de caractéristiques de haute dimension sur un espace 3D afin de faciliter la visualisation et permet ainsi une interaction intuitive de l'utilisateur. En forçant la position de certains documents agents dans l'espace, les agents deviennent automatiquement des agents requête, ne laissant d'autres choix aux autres agents libres que de réarranger leurs positions autour du ou des agents fixés.

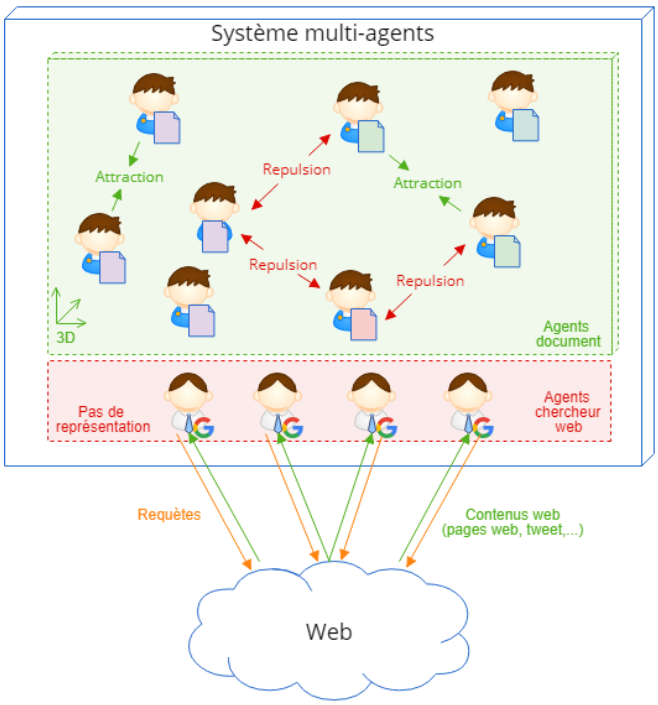



**Figure 8** Les agents document interagissent les uns avec les autres grâce à des actions de type attraction/répulsion. Les agents de recherche construisent des requêtes à partir des mots associés au *signal faible*.

## 6. Conclusion

Nous proposons une approche de recherche de thèmes communs dans un corpus de documents et la détection d'un sujet lié à un *signal faible* caractérisé par un petit nombre de mots par document et présent dans peu de documents. La combinaison LDA/Word2Vec tel que nous avons proposé de le mettre en œuvre nous permet de nous libérer du choix arbitraire du paramètre K (nombre de clusters) pendant le partitionnement. Deux directions étaient explorées : 1) l'algorithme 1 vise à trouver le nombre de sujets conduisant à un partitionnement par LDA aussi cohérent que possible ; 2) l'algorithme 2 qui, d'une manière plus avancée, combine les meilleurs sujets retournés par LDA sur l'ensemble de l'arborescence construite lorsque K varie.

Dans le cadre de cette étude sur la détection des *signaux faibles* et l'émission d'alertes, le système proposé cherche à mettre en corrélation les informations transmises initialement avec un contexte informationnel plus large grâce à des phases

d'exploration sur les réseaux. L'utilisateur interagit avec le système multi-agents pour guider les requêtes sur le web.

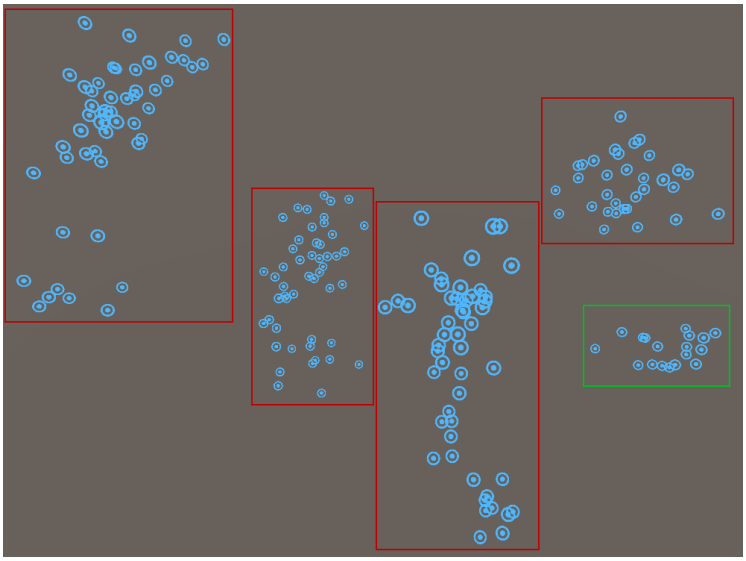


**Figure 9** Notre système multi-agents en action démontre sa capacité à auto-organiser les documents dans un espace 3D mais aussi à faire émerger des clusters. Les cases rouges représentent quelques thèmes principaux et la case verte représente le thème "*signal faible*".

Kim J, Lee C. Novelty-focused weak signal detection in futuristic data: Assessing the rarity and paradigm unrelatedness of signals. Technol Forecast Soc Change 2017;120(June 2016):59-76.

Park C, Cho S. Future sign detection in smart grids through text mining. Energy Procedia 2017;128:79-85.

Alghamdi R, Alfalqi K. A Survey of Topic Modeling in Text Mining. IJACSA) Int J Adv Comput Sci Appl 2015;6(1):147-53.

Levy O, Goldberg Y. Neural Word Embedding as Implicit Matrix Factorization. Adv Neural Inf Process Syst 2014a:.

Levy O, Goldberg Y. Linguistic Regularities in Sparse and Explicit Word Representations. 2014b.

Mikolov T, Corrado G, Chen K, Dean J. Efficient Estimation of Word Representations in Vector Space. Proc Int Conf Learn Represent (ICLR 2013) 2013:1-12.

Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. Distributed Representations of Words and Phrases and their Compositionality. CrossRef List Deleted DOIs 2013:1-9.

Socher R, Bauer J, Manning CD, Ng AY. Parsing with compositional vector grammars. ACL 2013 - 51st Annu Meet Assoc Comput Linguist Proc Conf 2013:

Socher R, Perelygin A, Wu J. Recursive deep models for semantic compositionality over a sentiment treebank. Proc 2013 Conf Empir methods Nat Lang Process 2013:1631-42.

Thorleuchter D, Van Den Poel D. Weak signal identification with semantic web mining. Expert Syst Appl 2013;40(12):4978-85.

Yoon J. Detecting weak signals for long-term business opportunities using text mining of Web news. Expert Syst Appl 2012;39(16):12543-50.

Hiltunen E. The future sign and its three dimensions. Futures 2008;40(3):247-60.

Nallapati RM, Ahmed A, Xing EP, Cohen WW. Joint latent topic models for text and citations. Proc 14th ACM SIGKDD Int Conf Knowl Discov data Min 2008:

Shen Z-Y, Zhi-Yong S, Sun J, Jun S, Yi-Dong S, Shen Y-D. Collective Latent Dirichlet Allocation. 2008.

Globerson A, Chechik G, Pereira F, Tishby PN. Euclidean Embedding of Co-occurrence Data. J Mach Learn Res 2007;8:2265-95.

Blei DM, Lafferty JD. Dynamic topic models. 2006.

Rigouste L, Cappé O, Yvon F. Quelques observations sur le modèle LDA. Journées Int d'Analyse Stat des Données Textuelles 2006;8.

Day GS, Schoemaker PJH. Scanning the Periphery. Harv Bus Rev 2005;83(11):135-48.

Decker R, Wagner R, Scholz SW. An internet-based approach to environmental scanning in marketing planning. Mark Intell Plan 2005;23(2):189-99.

Blei DM, Ng AY, Jordan MI. Latent Dirichlet Allocation. J Mach Learn Res 2003;3:993-1022.

Hofmann T. Unsupervised learning by probabilistic Latent Semantic Analysis. Mach Learn 2001;42(1-2):177-96.

Coffman B. Weak signal research, part I: Introduction. J Transit Manag 1997:

Ansoff HI. Managing Strategic Surprise by Response to Weak Signals. Calif Manage Rev 1975;XVIII(2):21-33.