Détection de signaux faibles dans des masses de données faiblement structurées

Detection of weak signals in weakly structured data masses

Julien Maitre[1], Michel Menard[1], Guillaume Chiron[1], Alain Bouju[1]

[1] L3i - Univ. La Rochelle, Avenue Michel Crépeau, La Rochelle, France
{julien.maitre,michel.menard,guillaume.chiron,alain.bouju}@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. Cet article se focalise principalement sur la recherche de *signaux faibles* présents dans les documents. Il s'agit d'une problématique investiguée dans un grand nombre de champs disciplinaires et de cadres applicatifs. Nous supposons que chaque document est un mélange d'un petit nombre de thèmes ou catégories, et que la création de chaque mot est attribuable en termes de probabilités à l'un des thèmes du document. Les catégories des documents transmis ne sont pas connues *a priori*. Les mots-clés présents dans les documents représentatifs de ces catégories sont également inconnus. L'analyse des documents reçus doit simultanément permettre de découvrir les thèmes, classer les documents relativement à ces thèmes, détecter les mots-clés pertinents relatifs aux thèmes et enfin découvrir les mots-clés relevant d'un thème "signal faible" éventuel. Pour atteindre cet objectif, nous proposons une définition du signal faible qui conditionne l'approche conjointe modèle thématique / plongement lexical, et contraint le choix des méthodes *LDA* et *Word2Vec*. Nous proposons d'évaluer les partitions obtenues grâce à un indice de cohérence sur la collection de mots représentative de chaque thème obtenu. Les clusters obtenus sont ainsi plus cohérents au sens contextuel. La détection du cluster associé au signal faible est alors plus aisée et plus pertinente.

ABSTRACT. This paper is related to a project aiming at discovering weak signals from different streams of information, possibly sent by whistleblowers in a platform as GlobalLeaks. 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.

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. Être capable de détecter rapidement 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 1.1), 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 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.

D'une manière générale, nous inscrivons cette étude dans le cadre des data journalistes qui reçoivent de la part de

lanceurs d'alerte des masses de documents (courriers électroniques, notes et rapports internes, documentation, . . .). Au-delà d'un simple stockage de l'information, les outils d'aide à l'investigation doivent pouvoir traiter, analyser et hiérarchiser cette information hétérogène : identifier les thèmes présents dans ces documents (relatifs par exemple à des évènements ou à des centres d'intérêt de communautés) et les mots-clés présents dans les documents associés à ces thèmes. Le journaliste doit pouvoir se servir ensuite de cette information structurée pour poursuivre son investigation en ayant recours à d'autres médias, et ainsi évaluer les corrélations et les enjeux. Pour anticiper les évènements, il doit nécessairement identifier les *signaux faibles* cachés dans la masse d'information. Il s'agit donc d'une analyse quantitative à forte valeur ajoutée (smart data).

Un des premiers exemples les plus marquants de data journalisme a porté sur l'étude des documents relatifs à la guerre d'Irak et d'Afghanistan divulgués par Chelsea Mannin via la plateforme WikiLeaks. Les premiers documents diffusés ont été des rapports de terrain. Sur la plateforme WikiLeaks, 391 000 rapports sont actuellement organisés par type, catégorie, date, et mot-clé. Des patterns (actuellement phrases et mots-clés) permettent de relier certains documents entre eux. Ces patterns sont détectés manuellement grâce à des experts. Comme il est précisé sur le site, ce sont ces patterns qui ont permis de créer ces liens (`https://wardiaries.wikileaks.org/`) au départ invisible. Dans notre étude, nous faisons l'hypothèse qu'il existe des patterns caractéristiques de *signaux faibles* (non encore identifiés à ce jour) qui permettent de détecter d'autres corrélations entre documents.

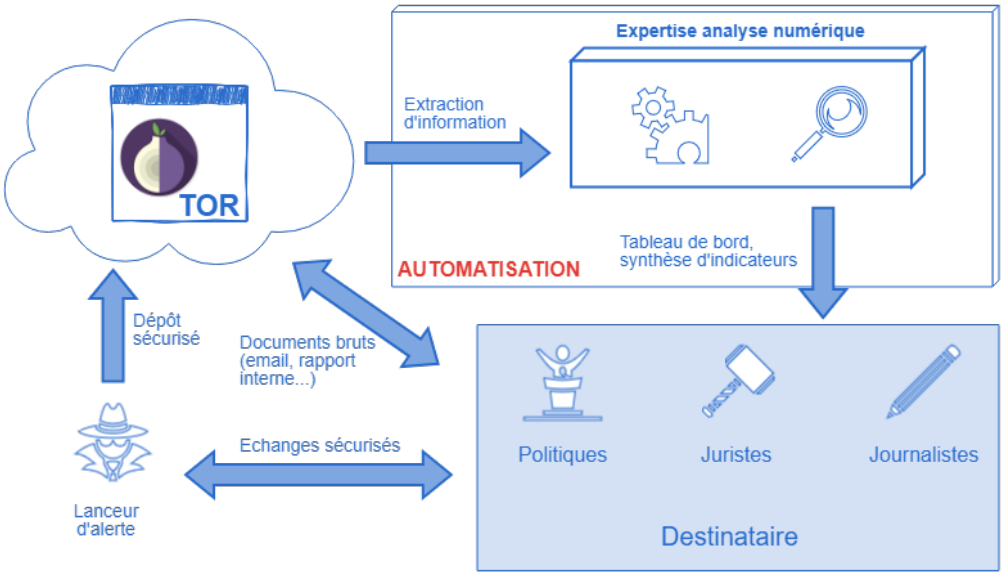

FIGURE *1.1.* *Aperçu de notre plateforme d'investigation. Elle doit répondre au besoin réel des journalistes/politiciens/juristes de disposer d'outils d'investigation (extraction, vérification, corrélation) et de représentation de l'information (synthèse, aide à la décision). Son but est donc de faciliter 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.*

Le projet vise à établir une procédure d'enquête capable d'aborder les actions suivantes. **A1** : Analyse automatique de contenus avec un minimum d'*a priori*. Identification des informations pertinentes. Calcul d'indicateurs de cohérence des thèmes identifés. Détection des *signaux faibles*. **A2** : Agrégation de connaissances. Enrichissement de l'information. **A3** : Visualisation analytique. Mise en perspective de l'information par la création de représentations visuelles et de tableaux de bord dynamique.

Plus précisément, les contributions décrites ci-après portent essentiellement sur ces actions et proposent respectivement une solution pour (1) la détection des *signaux faibles*, (2) l'extraction des informations qu'ils véhiculent et (3) la valorisation des informations de manière interactive. Notre système extrait, analyse et met automatiquement les informations dans des tableaux de bord. Il construit des indicateurs pour les destinataires qui peuvent également visualiser l'évolution dynamique de l'information gérée par un système multi-agent. Actuellement, plutôt que d'utiliser PCA ou tSNE (Van Der Maaten & Hinton, 2008) pour visualiser nos documents dans un espace 2D

réduit, nous avons opté pour un système multi-agents "d'attraction/répulsion" dans lequel les distances entre agents (i.e. documents) sont déterminées par leurs similarités (concernant leurs caractéristiques extraites). Cette approche a l'avantage d'offrir à la fois des capacités d'évolution en temps réel et une riche interaction pour l'utilisateur final (par exemple, en forçant la position de certains agents).

La Figure 1.1 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.

L'article est organisé comme suit : tout d'abord nous présentons un premier état de l'art sur cette problématique afin d'éclairer le contexte de l'étude et 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*. Ensuite, un état de l'art plus technique est fourni pour la modélisation des clusters et les méthodes de plongement de mots qui sont toutes deux impliquées dans la solution que nous proposons. La section 4 présente l'une de nos contributions : une solution "*LDA* [1] augmentée avec *Word2Vec*" pour la détection des *signaux faibles*. Quelques résultats dans un premier temps sur un corpus artificiel, puis un second temps sur un corpus de données réelles, montrent en section 5 l'intérêt de cette approche ainsi qu'une proposition de visualisation interactive permettant de gérer en temps réel les documents porteurs de *signaux faibles*.

2 Etat de l'art sur les *signaux faibles*

2.1 Nature des signaux faibles

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 & Schoemaker, 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 & Lee, 2017; Yoon, 2012).

Pour surmonter ces limites, Yoon (Yoon, 2012) a proposé une approche s'appuyant sur une carte d'émergence de

1. Latent Dirichlet Allocation

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 & Cho, 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 & Lee, 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. Comme le décrit Ah-Pine (Ah-Pine *et al.*, 2005) lors du développement de « RARES Text », la plausibilité qu'une information relative à une classe de documents soit un *signal faible* est d'autant plus grande que celle-ci s'agrège « difficilement » dans les niveaux hiérarchiques supérieurs. Thorleuchter (Thorleuchter & Van Den Poel, 2013) complètent cette définition par le fait que les mots-clés du *signal faible* sont sémantiquement reliés. Il ajoute donc le qualificatif de dépendance.

Tout comme la détection de *signaux faibles*, la détection de nouveauté est une tâche d'apprentissage non supervisée qui vise à identifier des échantillons inconnus ou incohérents d'un ensemble de données. Chaque approche développée dans la littérature se spécialise dans une application particulière telle que le diagnostic médical (Clifton *et al.*, 2011), la surveillance de systèmes industriels (Ebrahimkhanlou & Salamone, 2017) ou le traitement vidéo (Ramezani *et al.*, 2008). (Mohammadi-Ghazi *et al.*, 2018; Domingues *et al.*, 2018) présentent plusieurs références concernant la détection de nouveauté. Dans l'analyse de document, les approches les plus utilisées et les plus pertinentes sont construites en utilisant l'allocation de Dirichlet latente (*LDA*). Dans cet article, nous avons choisi de comparer notre approche à *LDA*.

2.2 Positionnement

Les qualificatifs des *signaux faibles* ou des mots-clés associés que nous souhaitons détectés sont cohérents avec ceux que l'on retrouve dans la littérature même si utilisés dans d'autres contextes : unitaire, rareté, non relié à des paradigmes existants, nouveauté, anormalité, sémantiquement reliés (Kim & Lee, 2017; Thorleuchter & Van Den Poel, 2013; Ah-Pine *et al.*, 2005). La définition que nous adoptons, mise en œuvre par l'utilisation conjointe des approches "topic modeling" et "word embedding", permet dans l'arborescence des clusters/thèmes découverts de détecter les plus cohérents au sens de la notion de dépendance entre mots-clés : des groupes de mots-clés apparaissent conjointement dans les documents, ils doivent appartenir à un seul et même cluster fortement cohérent (unitaire, nouveauté) et disjoint des autres (donc non relié sémantiquement à d'autres thèmes). Le nombre d'occurrence de ces mots-clés est également plus faible et ces mots-clés sont présents dans peu de documents (rareté).

Nous retenons donc pour notre étude la définition suivante des *signaux faibles* :

Definition. 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).

Nous supposons que chaque document est un mélange d'un petit nombre de thèmes ou de catégories, et que la création de chaque mot est attribuable en termes de probabilités à l'un des thèmes du document. Un *signal faible* correspond à un thème spécifique fortement cohérent. L'approche méthodologique proposée doit permettre de l'identifier/le révéler (comme pour les autres thèmes) à partir de sa collection de mots présents dans les documents grâce à un modèle génératif probabiliste permettant d'expliquer les ensembles d'observations/documents.

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 2.2 illustre la chaîne de traitement. Nous nous focalisons dans cet article sur le problème du clustering multi-niveaux, et nous présentons notre approche multi-agents construit sur un schéma d'attraction/répulsion.

Nous ne prenons pas en compte l'aspect temporel qui pour être exploité, requiert un corpus de documents datés, ce qui n'est pas toujours le cas. Dans le cas où des dates sont disponibles, il n'est d'ailleurs pas garantie qu'elles soit fiables, en particulier si elles sont issues d'un processus d'extraction automatique. C'est pourquoi, nous préférons écarter dans un premier temps une approche qui s'appuie fortement sur une chronologie éventuelle des documents.

Dans cette étude, nous prenons donc comme hypothèse que le *signal faible* émane d'une information partielle et fragmentaire relative à un fait particulier agissant comme révélateur. Les mots–clés portant et décrivant cette information sont ainsi plus resserrés. Par exemple si on se réfère au scandale des boues rouges déversées en mer Méditerranée, les premiers articles sur le web qui y ont fait référence (parmi d'autres articles publiés par un agrégateur de contenus), se sont focalisés sur un fait substantiel (e.g. description d'un acte de pollution localisé) et ont utilisé un descriptif cohérent, resserré et spécifique au sens sémantique, qui peut être qualifié de pattern textuel. Généralement un journaliste se sert ensuite de ce pattern pour identifier des faits se rapportant au même scandale dans d'autres sources d'informations. Il est naturel de considérer ce pattern (présentant donc une cohérence forte et dont les mots-clés sont particulièrement liés) comme suffisamment discriminant, et donc disjoint des autres thèmes dont le vocabulaire descriptif est nécessairement moins resserré (i.e. ne correspondant pas à un pattern textuel). Ceci rejoint les travaux de Ah-Pine (Ah-Pine *et al.*, 2005) qui décrit notamment une information nouvelle comme étant relativement orthogonale aux autres informations contenues dans le corpus.

3 Modélisation thématique et exploration d'une collection de documents

Cette section apporte aux lecteurs un état de l'art des méthodes de "modélisation thématique" et de "plongement de mots" qui sont des méthodes de *traitement automatique des langues* (*TAL*). Elles appartiennent à des paradigmes différents mais sont complémentaires dans la solution que nous proposons.

3.1 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 & Alfalqi, 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 & Lafferty, 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 (Deerwester *et al.*, 1990; Hofmann, 2001; Alghamdi & Alfalqi, 2015).

LSA apporte une représentation de faible dimension des documents et des mots. Elle s'appuie sur une matrice qui décrit l'occurrence du mot dans le document (nombre brut ou normalisée par *tf-idf*). On utilise ensuite une décomposition en valeurs singulières permettant, outre la réduction de la dimension, de traduire les vecteurs mots

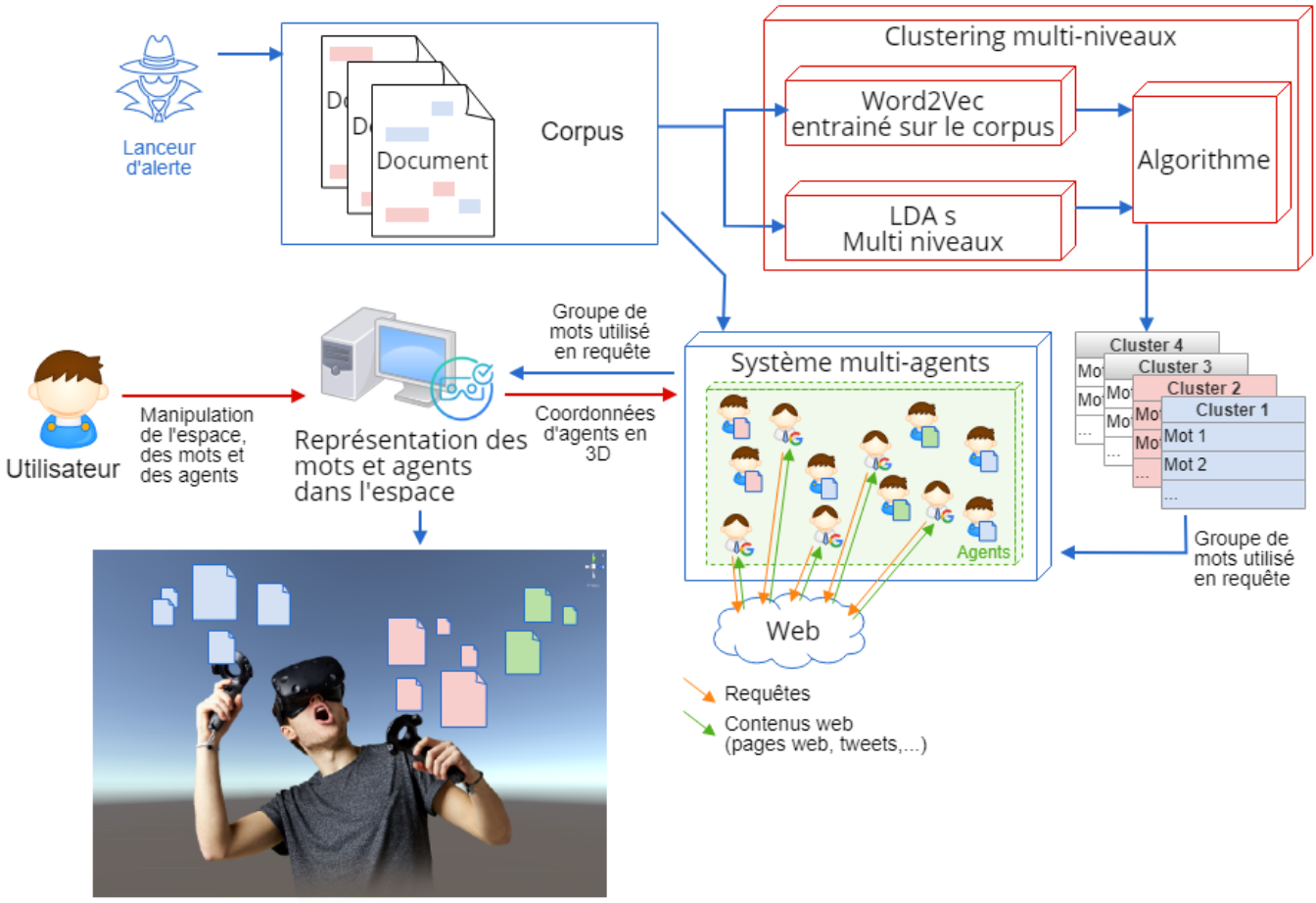

FIGURE 2.2. *Le système extrait et analyse automatiquement les informations fournies par le lanceur d'alerte. Il construit des indicateurs placés ensuite dans des tableaux de bord pour analyse et consultation. Ceux-ci sont présentés de manière dynamique grâce à un système multi-agents dans un objectif de navigation et de recherche documentaire. Dans ce but, chaque document est représenté par un agent se déplaçant dans un environnement 3D.*

et documents dans l'espace des concepts (les concepts sont supposés être orthogonaux). Il est ainsi possible de relier les documents entre eux (Deerwester *et al.*, 1990).

Afin de mieux faire correspondre le modèle statistique aux données observées, une approche de nature probabiliste a été développée reposant sur des distributions multinomiales, l'analyse sémantique latente probabiliste, *PLSA*. Les concepts/thèmes ne sont plus contraints à être orthogonaux.

Cependant, la phase d'identification des paramètres augmente linéairement avec le nombre de documents, le système devient de plus en plus complexe. Par conséquent *PLSA* souffre d'un sur-apprentissage. Comme l'algorithme est itératif et converge lentement, le temps d'exécution augmente fortement, spécialement avec les grands jeux de données (Alghamdi & Alfalqi, 2015; Kakkonen *et al.*, 2008) et la distribution des thèmes reste concentrée sur un nombre limité de sujets (Hofmann, 2001).

Au cours des années 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).

Dans le cadre du text mining, *LDA* est un algorithme d'exploration largement utilisé s'appuyant sur un processus de Dirichlet (statistiques bayésiennes). Il peut servir de modèle génératif pour l'imitation du processus d'écriture (Alghamdi & Alfalqi, 2015; Kakkonen *et al.*, 2008). 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).

3.2 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 des langues* (*TAL*) 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 (Bakarov, 2018).

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

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 loin dans le texte *w2vSim*). Une valeur proche de 1 indique que deux mots sont très proches l'un de l'autre (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 *TAL* telles que l'analyse syntaxique (Köhn, 2016; Bansal *et al.*, 2014; Socher *et al.*, 2013b), la détection de métaphore (Tsvetkov *et al.*, 2014, 2015), la reconnaissance d'entités nommées (Turian *et al.*, 2010; Collobert *et al.*, 2011), l'analyse des sentiments (Schnabel *et al.*, 2015; Socher *et al.*, 2013a) et la détection de paraphrase (Baumel *et al.*, 2016; Bakarov & Gureenkova, 2018).

3.3 Justification de l'approche LDA pour l'extraction des signaux faibles

A la différence de PLSA, l'approche *LDA* utilise une distribution de Dirichlet, ce qui évite le sur-apprentissage et favorise la dispersion des documents sur de nombreux thèmes différents.

De plus *LDA* est un modèle génératif et se généralise très bien à de nouveaux documents non présents dans l'ensemble initial. Cela simplifie la tâche difficile qui est l'ajout de nouveaux documents au modèle lors du processus d'estimation (Kakkonen *et al.*, 2008).

Ceci nous paraît essentiel pour permettre la détermination d'un cluster représentatif du *signal faible* de part ses propriétés de nouveauté et d'anormalité. En effet, la presse [2] publie généralement les révélations en plusieurs fois. La collection de documents n'est donc pas fixe. Il est donc crucial d'avoir un modèle suffisamment flexible pour gérer correctement un document qui n'a pas été vu auparavant.

Nous ne présupposons pas que les thèmes supportés par les documents puissent être décrits d'une manière hiérarchique, ce que laisserait supposer l'utilisation de *hLDA* [4]. Puisque l'objectif est la détection d'un cluster relatif au *signal faible*, celui-ci est par définition disjoint des autres (unitaire et non relié sémantiquement aux autres clusters i.e. à des paradigmes existants), et selon la définition de Ah-Pine (Ah-Pine *et al.*, 2005) relativement orthogonale aux autres informations contenus dans le corpus.

2. Par exemple, 70 000 documents confidentiels sur les opérations de la coalition internationale en Afghanistan ont été diffusés par le site WikiLeaks [3] en juillet 2010, puis 400 000 rapports concernant l'invasion américaine en Irak sont ensuite publiés en octobre et enfin le contenu de 250 000 câbles diplomatiques. D'autres exemples peuvent être également cités.

4. Ce dernier est un modèle générique de hiérarchie de clusters où chaque document est généré selon un chemin partant de la racine jusqu'à une feuille en échantillonnant les sujets le long de ce chemin et en échantillonnant les mots des sujets sélectionnés (Blei *et al.*, 2004). Ainsi les clusters sont tous liés à travers l'arbre.

3.4 Justification d'une approche conjointe modèle thématique/plongement lexical

Toutes les approches de clustering, qu'elles soient de nature heuristiques, possibilistes, probabilistes, ou floues (construites par exemple à partir de fonctions objectives) souffrent de la même difficulté : le nombre de clusters obtenus ne correspond en général qu'à un optimum local. En effet, le problème peut avoir plusieurs solutions. Le choix du critère s'avère être donc difficile car il suppose qu'on ait une bonne définition de ce qu'est un cluster qui peut être de forme et de taille quelconques. De plus, les données du problème sont entachées de bruits (outliers) et d'ambiguïté entre clusters (Gunes *et al.*, 2010; Ménard, 2001; Ménard & Eboueya, 2002). Même si les approches deviennent de plus en plus robustes à ces artefacts, notamment grâce aux méthodes construites autour des processus de Dirichlet, la détermination du nombre de clusters reste sensible à la structuration des observations et à l'information *a priori* disponible. La famille des algorithmes *LDA* n'échappe pas à cette difficulté et il est souvent proposé une mise en œuvre de l'algorithme avec un nombre de clusters recherchés très important, suivi d'une évaluation heuristique, afin d'agréger les clusters. Un nombre excessif de clusters peut conduire à un modèle trop complexe à évaluer sans l'aide d'experts. D'autres approches proposent d'effectuer plusieurs tests avec un paramétrage différent et utilisent des critères de validation croisée. Des critères heuristiques existent cependant, construits par exemple sur l'erreur quadratique, les matrices de covariance ou encore sur la notion de perplexité. Pour pallier au problème de stabilité dans le contexte topic modeling, (Zhao *et al.*, 2015) a proposé un critère informationnel s'appuyant sur le taux de changement de perplexité.

Dans le cadre de la résolution d'un problème inverse, un critère global pourra être construit afin d'assurer un compromis entre un terme d'attache aux données et un terme de régularisation de la solution. C'est le cas par exemple des approches de clustering de type possibilistes ou floues qui sont construites à partir de fonctions objectives. Ces critères doivent être construits en prenant en compte la structuration des données observées et sur les propriétés attendues des clusters recherchés. Nous proposons dans cette étude une approche méthodologique similaire afin de réaliser un compromis entre l'exploration d'une collection de documents et une représentation interne de séquences de mots, reposant pour la première sur la modélisation thématique, et pour la seconde sur le plongement lexical.

Concernant le critère, l'objectif de l'étude est la détermination d'un *signal faible* représenté par un cluster de mots, nous proposons donc de s'appuyer sur la définition proposée du *signal faible* afin de le mettre en évidence.

Nous prônons l'utilisation d'une approche conjointe modèle thématique et plongement lexical. La première vise principalement à décrire des documents et des collections de documents en leur assignant des distributions de thèmes, qui à leur tour ont des distributions de mots assignés. Elle capture ainsi des associations au niveau des documents. La seconde cherche à positionner des mots dans un espace vectoriel latent. Elle n'est pas vraiment conçue pour décrire des documents mais permet la capture des associations très locales. L'approche conjointe que nous proposons repose sur l'utilisation de *LDA* standard et de *Word2Vec*.

Pour *Word2Vec*, les mots sont représentés par un vecteur de longueur fixe et attribue une signification sémantique aux distances entre les représentations des mots, ce qui est une des caractéristiques recherchées des signaux faibles (mots sémantiquement reliés). Les deux approches s'avèrent donc complémentaires car le modèle s'applique pour la première sur la représentation d'un document par un vecteur de longueur fixe, le second s'attache à décrire un mot par un vecteur de longueur fixe.

Dans un premier temps, pour plusieurs valeurs du nombre de clusters, nous appliquons *LDA* standard sur le corpus de documents. Puis dans un deuxième temps, grâce à un indicateur de ressemblance nous construisons une arborescence de clusters. Cette arborescence est finalement simplifiée et élaguée grâce au critère de cohérence, et seuls les clusters cohérents au sens de notre définition des signaux faibles sont alors retenus.

Il est important de remarquer que *Word2Vec* doit calculer les poids du réseau sur un corpus de documents. Un espace de représentation associé et dépendant du corpus est alors calculé. Un modèle entrainé sur un large corpus de documents traitant de plusieurs domaines sera plus générique dans la représentation de ses mots par rapport à un modèle traitant d'un corpus portant sur un domaine spécifique.

Il est important de remarquer que *Word2Vec* doit calculer les poids du réseau sur un corpus de documents. Un espace de représentation associé et dépendant du corpus est alors calculé. Un modèle entrainé sur un large corpus de documents traitant de plusieurs domaines sera plus générique dans la représentation de ses mots, rendant moins

pertinent le plongement sur le problème rencontré, par rapport à un modèle traitant d'un corpus portant sur un domaine spécifique. L'avantage de définir *Word2Vec* sur un corpus limité (voir 5.1 dans la section Expérimentation) est d'obtenir un plongement spécifique au corpus et plus performant. Dans le cas des données extraits du corpus Wikipédia (voir 5.2), nous utiliserons un réseau disponible en ligne et appris sur l'encyclopédie Wikipédia. Il est cependant tout à fait possible de limiter le corpus aux documents transmis par le lanceur d'alerte afin de renforcer la spécificité du critère contextuel. Il sera alors nécessaire lors de la transmission d'autres documents par ce même lanceur d'alertes d'entrainer sur la totalité des documents afin de faire bénéficier d'un même espace de représentation tous les mots du corpus.

4 *LDA* augmenté avec *Word2Vec*

Cette section décrit notre contribution consistant en l'utilisation d'un critère construit à partir de *Word2Vec* pour filtrer/sélectionner les clusters les plus cohérents parmi ceux fournis par *LDA* et gérés à différents niveaux K. Afin de faciliter la compréhension de l'approche, la figure 4.3 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.

LDA est une méthode de clustering (non supervisée) et n'associe pas un label aux clusters trouvés. Dans le cas d'un grand nombre de documents, il est difficile d'estimer *a priori* le nombre de clusters potentiels. De plus, il est difficile de discerner la cohérence de chaque groupe. Pour cela, il est nécessaire de définir un indicateur, c'est l'objet de la section suivante.

Simplement à titre d'illustration, le tableau 4.1 donne un exemple de ce que *LDA* permet d'obtenir pour un corpus de données réelles : le mélange de thèmes présents dans chaque document et les mots associés à chaque thème. Pour détecter un cluster relatif au *signal faible* selon la définition choisie, il est nécessaire d'évaluer la cohérence de chaque thème. Ceci s'effectue grâce à une méthode de plongement de mots.

	Thème 1	Thème 2	Thème 3	Thème 4
Mots	commune	film	saison	guerre
	ville	album	club	pays
	roi	premier	première	france
	nom	années	premier	général
	église	ans	tour	français
	...	...	...	...

Tableau 4.1. *Liste des 5 premiers mots de chaque thème*

4.1 Indicateur de cohérence en tant que mesure intra-cluster

Le premier indicateur de cohérence locale proposé repose sur la méthode de plongement de mots, *Word2Vec*. Il propose de qualifier la similitude sémantique intrinsèque d'un ensemble de mots (clusters) dans le contexte du corpus de documents. Ce premier indicateur est défini comme suit :

$$I_1 = \sum_{w \epsilon E} w2vSim(w_i, w_j) \qquad [1]$$

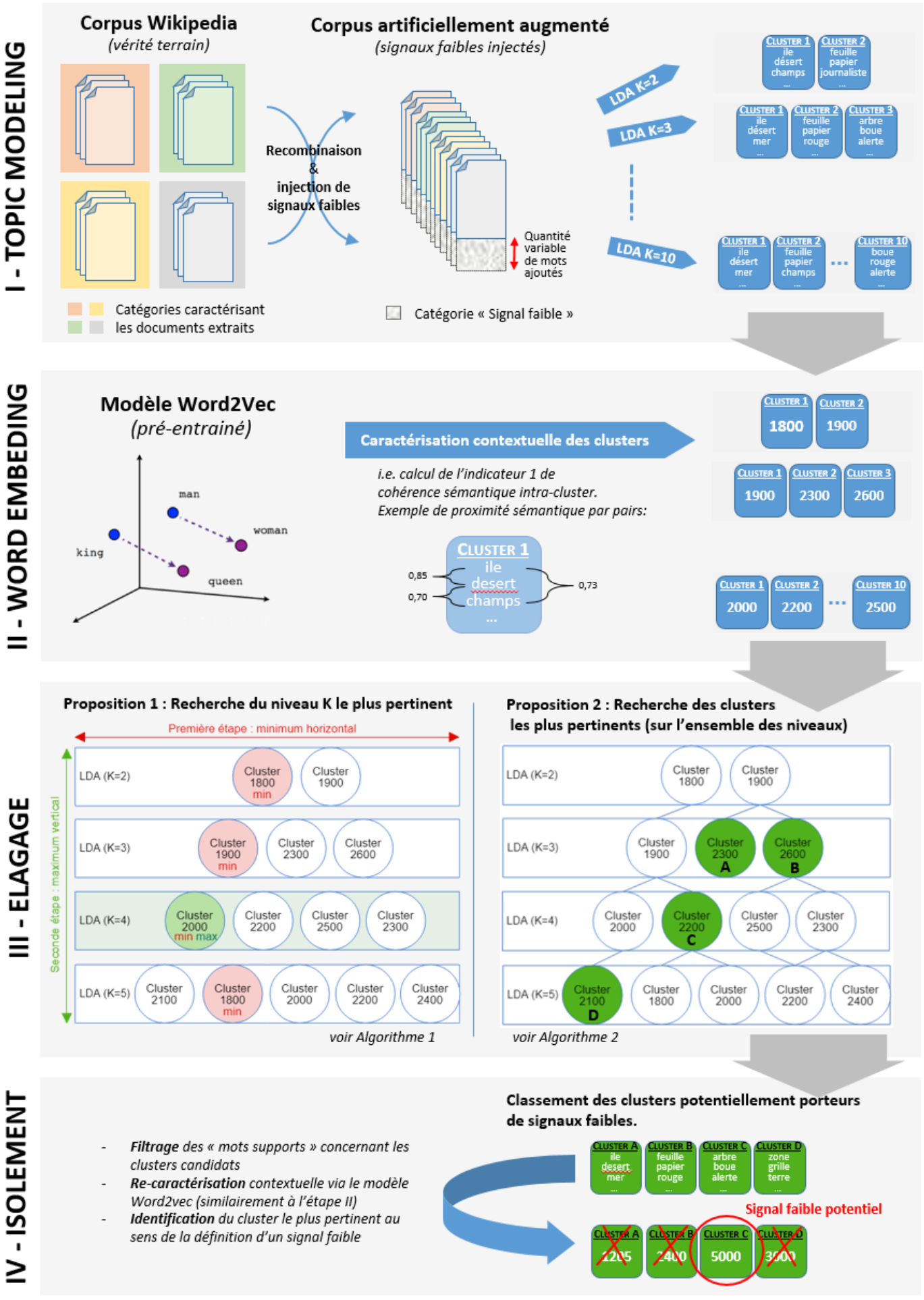

FIGURE 4.3. *Les différentes étapes pour extraire les signaux faibles*

où I_1 représente la somme des valeurs de similarité de toutes les combinaisons de paires de mots dans chaque cluster, avec $E = \{w_1, w_2, \ldots, w_{100}\}$ l'ensemble des 100 premiers mots supportant le cluster. $w2vSim$ représente la mesure de similarité définie dans *Word2Vec* qui est la similarité cosinus entre deux vecteurs (Mikolov *et al.*, 2013a). Plus la valeur est grande, plus le cluster contient de mots régulièrement employés ensemble. Le choix arbitraire des 100 premiers mots s'est avéré être un bon compromis : il permet à la fois de constituer une liste courte de mots clés relative à un thème, et une interprétation encore aisée par un expert des résultats obtenus. Les mots sont pondérés par leur probabilité. Ce choix a été ramené à 10 mots sur les tests effectués sur les corpus proche du réel.

Nous proposons d'utiliser cet indicateur sur les clusters découverts par l'algorithme *LDA*. Plusieurs *LDA* sont appliqués avec différentes valeurs de k. Nous obtenons ainsi plusieurs partitions qui peuvent alors être représentées sous la forme d'une arborescence. Il est à noter que *LDA* organise les clusters découverts lors des différentes éxécutions dans un ordre aléatoire. Une étape supplémentaire présentée dans la section 4.3 est donc nécessaire pour construire cette arborescence.

Afin d'évaluer les partitions obtenues, deux algorithmes sont proposés à partir de l'indicateur précédent : 1) un premier algorithme (décrit dans la Section 4.2) visant à rechercher le nombre de clusters (paramètre k) conduisant à un partitionnement par *LDA* le plus cohérent possible vis-à-vis de cet indicateur ; 2) un algorithme (décrit dans la Section 4.3) qui, de manière plus avancée, par une analyse approfondie de l'arborescence, combine les meilleurs clusters renvoyés par *LDA* sur toutes les partitions (ou valeurs de k testées).

4.2 Recherche du paramètre k conduisant aux clusters les plus pertinents au sens du critère de cohérence

L'algorithme 1 consiste à rechercher le niveau de l'arborescence donnant les clusters les plus cohérents au sens de l'indicateur I_1. A chaque niveau, nous calculons la valeur minimale de cet indicateur parmi ceux calculés sur l'ensemble des clusters présents sur ce niveau. Le niveau k choisi (et donc le nombre de clusters pertinents au sens du critère) correspond à celui dont la valeur minimale est la plus élevée (cf. Eq. 2 et Figure 4.3).

$$\underset{k}{Argmax}(min_{c_l}(I_1(c_l))), c_l \in \{c_1 \ldots c_k\}, k \in \{2 \ldots K\} \quad [2]$$

Data: P = Liste des nombres de clusters demandés : {2...K}
Result: meilleurk = identifiant du k niveau
meilleurk $\leftarrow 2$;
meilleurScorek $\leftarrow min_{c_l}(I_1(c_{meilleurk}))$;
for $k \in P$ **do**
  for $c_l \in \{c_1 \ldots c_k\}$ **do**
    if $min_{c_l}(I_1(c_l)) >$ meilleurScorek **then**
      meilleurk $\leftarrow k$;
      meilleurScorek $\leftarrow min_{c_l}(I_1(c_l))$;
    end
  end
end
return meilleurk

Algorithm 1: Recherche du niveau de l'arborescence donnant les clusters les plus cohérents

4.3 Une approche heuristique pour déterminer les clusters les plus pertinents sur l'ensemble de l'arborescence LDA

Pour construire l'arborescence, il est nécessaire d'évaluer un lien de ressemblance entre les clusters de différents niveaux. Celui-ci est calculé en utilisant un indicateur de ressemblance utilisant la distance de Bhattacharyya

définie comme suit (Bhattacharyya, 1943) :

$$I_2 = \sum_{w \epsilon E} \sqrt{p_{w_i} \cdot q_{w_i}} \quad [3]$$

Pour l'ensemble E défini par les mots w présents dans un cluster c_l de niveau k et un cluster c_{l+1} de niveau $k+1$, nous calculons la somme des produits de probabilités, p_{w_i} et q_{w_i}, de chaque mot présent dans les clusters respectifs c_l et c_{l+1}.

Il est alors possible d'extraire sur toute l'arborescence les clusters les plus pertinents au sens du critère de cohérence donné par l'indicateur I_1, et des relations de similarité, indicateur I_2, entre deux clusters de niveau respectif k et $k+1$. Pour ce faire, nous proposons de parcourir l'arbre de façon récursive en suivant une exploration ordonnée à partir de l'indicateur I_2. Au cours de ce processus, chaque cluster nouvellement rencontré, retenu comme pertinent, conduit au retrait dans l'arborescence de tous les clusters parents et fils. Les relations entre les clusters (décrites par l'indicateur I_2) ne sont prises en compte qu'au-delà d'un seuil arbitrairement défini (cf. Figure 4.3). L'algorithme 2 formalise cette heuristique où les fonctions $Parents(c_l)$ et $Fils(c_l)$ récupèrent respectivement la liste des clusters parent et fils du cluster c_l. Nous obtenons alors une liste des clusters pertinents qui ne sont pas connectés au sens de l'indicateur I_2.

Afin d'évaluer la performance de l'approche, il est nécessaire de les confronter expérimentalement avec l'utilisation de *LDA* seul. Cette évaluation est abordée dans la section suivante.

Data: T = Liste des thèmes de l'arborescence LDA triés par valeur de cohérence
Result: themesRetenus = Liste des identifiants des thèmes pertinents
themesRetenus $\leftarrow$ {};
while `Taille(`T`)` > 0 **do**
  meilleurCluster $\leftarrow$ `Max(`T`)`;
  themesRetenus $\longleftarrow$ themesRetenus + {meilleurCluster };
  for $t \in$ `Parents(`meilleurCluster`)` **do**
    $T \longleftarrow T - t$;
  end
  for $t \in$ `Fils(`meilleurCluster`)` **do**
    $T \longleftarrow T - t$;
  end
end
return themesRetenus

Algorithm 2: Récupération des clusters pertinents dans l'arborescence *LDA*

L'algorithme 2 conduit aux propriétés suivantes :

$$\begin{cases} \overline{I_1(c_i)}^{Alg2} \geqslant \overline{I_1(c_i)}^{LDA_k} & \forall\ k \\ \max\limits_i I_1^{Alg2}(c_i) \geqslant \max\limits_i I_1^{LDA_k}(c_i) & \forall\ k \end{cases} \quad [4]$$

La moyenne des cohérences sémantiques des clusters trouvés est augmentée.

5 Experimentation

Une preuve de concept est réalisée pour évaluer l'approche présentée. A cette fin, nous avons créé 3 bases de données de documents : une base de données synthétiques utilisée comme vérité terrain dans l'expérimentation avec un corpus artificiel, une base de données sur des documents Wikipédia et une base de données sur des comptes rendus médicaux. Ces ensembles de données permettent une analyse objective et sans ambiguïté. Chaque corpus comportent plusieurs thèmes. Ils contiennent des documents considérés comme des mélanges de ces thèmes. Dans la base de données "synthétiques" et Wikipédia, les mots relatifs à un thème supplémentaire considéré comme

signal faible sont injectés dans tout ou partie des documents. Cela nous permet de déterminer plus objectivement la contribution des indicateurs I_1, I_2 et de l'algorithme 2 par rapport à *LDA*.

5.1 Test sur un corpus artificiel

Dans cette expérimentation, nous utilisons des corpus artificiels. Nous définissons les mots élémentaires utilisés dans le test pour construire différents corpus artificiels de documents. Ils sont composés de quatre thèmes principaux et d'un thème supplémentaire *signal faible*. Afin de mieux discerner le thème d'appartenance d'un mot, nous avons remplacé les mots par des nombres. Nous définissons ainsi des séries de nombres pour chaque thème :

— Thème 1 : nombres de 0 à 99
— Thème 2 : nombres de 100 à 199
— Thème 3 : nombres de 200 à 299
— Thème 4 : nombres de 300 à 399

Afin de tendre vers les conditions réelles, dans cette expérience, chaque document généré présente un thème principal auquel seront rattachés 10 000 mots, et deux thèmes secondaires représentés par 2 000 mots.

Un document texte est écrit avec des mots outils dont la syntaxe a préséance sur le rôle sémantique. Afin de tenir compte de cette caractéristique, nous ajoutons 60% de mots outils supplémentaires (par rapport aux 10 000 du document original). Un document textuel sera donc composé de 20 000 mots dérivés de 4 thèmes (3 thèmes principaux et le thème "*mots outils*") choisis parmi 5 thèmes.

	Nombre de mots du				
	Thème 1	Thème 2	Thème 3	Thème 4	Thème Mots-outils
Documents 1 à 50	10 000	2 000		2 000	6 000
Documents 51 à 100	2 000	10 000	2 000		6 000
Documents 101 à 150		2 000	10 000	2 000	6 000
Documents 151 à 200	2 000		2 000	10 000	6 000

Tableau 5.2. *Composition du corpus généré pour le test (en mots)*

Chaque document généré contient 20 000 mots appartenant aux 4 thèmes principaux ainsi qu'au thème "*mots outils*" comme décrit dans le tableau 5.2. Le sixième thème, présent dans une proportion variable de documents, est celui relatif au *signal faible*. Pour ce dernier, les mots sont distribués par groupes de 10 mots placés de manière aléatoire dans le document.

Un exemple d'un document généré est montré figure 5.4. Le thème associé au *signal faible* utilise des nombres entre 900 et 999. Dans l'expérience le cinquième thème représente les "*mots outils*" fréquents (e.g. le, et, ou). Ils correspondent à des nombres entre 600 et 699.

Nous présentons dans le tableau 5.3 un test, où dans un nombre variable de documents allant de 0 à 200 par incrément de 10, 600 mots du thème "*signal faible*" sont injectés et complètent les 20 000 mots provenant du thème principal, secondaires et relatif aux *mots outils*. Des tests additionnels avec 2, 4, 6 et 8 documents ont été ajoutés. Dans cette expérimentation proche des conditions réelles, où les documents sont multithématiques et composés de mots outils, les résultats obtenus montrent la robustesse de l'algorithme 2 même pour un très petit nombre de documents contenant des *signaux faibles* (3%).

L'identification du cluster *signal faible* s'effectue de la façon suivante :

— pour chaque cluster obtenu, nous calculons la somme des poids des mots n'appartenant qu'à un seul thème (e.g. les mots "Outils" ne sont pas pris en compte) ; Nous obtenons donc une valeur pour les thèmes 1 à 4 et 6 ;
— Le *signal faible* est détecté lorsque la valeur du thème 6 est la plus importante.

Ce test est effectué 10 fois en faisant varier le nombre de mots du signal faible injecté. (cf. Figure 5.3)

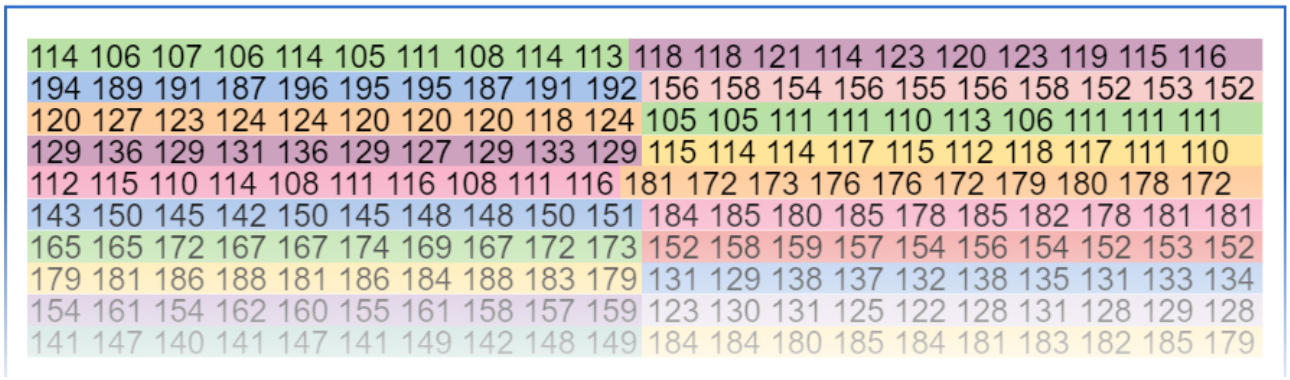

FIGURE 5.4. *Exemple de document. Les couleurs représentent des groupes de mots. Le document contient des mots du Thème* 2.

Documents avec signal faible présent	Algorithme 2	LDA						
		2	3	4	5	6	7	8
0	0	0	0	0	0	0	0	0
2	9	0	0	0	2	7	6	8
4	10	0	0	0	3	7	9	9
6	10	0	0	0	5	9	7	10
8	10	0	0	0	9	8	10	10
10	10	0	0	0	5	9	10	10
20	10	0	0	0	5	10	10	10
30	10	0	0	0	5	9	10	10
40	10	0	0	0	6	9	10	10
50	10	0	0	0	5	10	9	10
60	10	0	0	0	7	10	10	10
70	10	0	0	0	8	10	8	9
80	10	0	0	0	8	10	10	10
90	10	0	0	0	6	10	10	10
100	10	0	0	0	7	9	9	9
110	10	0	0	0	9	6	10	10
120	10	0	0	0	5	7	8	8
130	10	0	0	0	4	10	6	10
140	10	0	0	0	7	7	10	10
150	9	0	0	0	5	8	5	9
160	10	0	0	0	4	6	8	9
170	9	0	0	0	2	4	7	7
180	10	0	0	0	3	7	5	6
190	8	0	0	0	3	2	6	2
200	2	0	0	0	0	1	1	1
Total	227	0	0	0	123	185	194	207
Réussite	95%	0%	0%	0%	51%	77%	81%	86%

Tableau 5.3. *Test artificiel : résultat de l'application de l'algorithme 2 par rapport aux* 7 *LDA seuls paramétrés avec des valeurs de* k *allant de* 2 *à* 8*. Sur chaque ligne de la colonne "Documents avec signal faible présent" est donnée la proportion du nombre de documents portant le signal faible. Les résultats baissent (DocWithSecretWords = 200) quand tous les documents du corpus contiennent des mots relatifs au thème "signal faible" car il est alors considéré comme un thème de type "mots outils", et n'est donc pas détecté.*

Il n'est pas possible de s'assurer qu'une valeur de k donnée de *LDA* permet de détecter correctement le cluster relatif au *signal faible*. Même si il est détecté au niveau k, sa cohérence au sens du critère donné (défini dans la

section 4.1) peut être faible. De plus, pour une valeur donnée de k, *LDA* ne garantit pas l'observation des thèmes les plus pertinents. Certains groupes peuvent être cohérents à ce niveau de segmentation et d'autres non. Une décomposition trop profonde (i.e. une valeur de k trop grande pour ce jeu de données, e.g. *LDA* 8) conduit à un grand nombre de clusters qui ne sont pas nécessairement représentatives des thèmes de tous les documents (i.e. on assiste à une sursegmentation). Même si le cluster relatif au *signal faible* fait partie de l'un des clusters détectés, il peut quand même représenter seulement quelques mots du thème et non le thème dans son ensemble. L'intérêt de l'algorithme 2 est de balayer toute l'arborescence afin de détecter les clusters les plus cohérents (au sens du critère). Ces clusters peuvent être localisés à différents niveaux, ils ne sont plus limités à un seul niveau de l'arborescence. Parmi eux, celui correspondant au *signal faible* sera détecté à un niveau pertinent au sens de ce critère.

5.2 Test sur des corpus de données réelles

5.2.1 Test sur des documents Wikipedia

Concernant ce second test, nous présentons des résultats portant sur un sous-ensemble de documents extraits 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 réaliser le test, nous disposons d'un corpus de documents relatifs à 5 domaines :
— Economie : 44 876 documents
— Histoire : 92 041 documents
— Informatique : 25 408 documents
— Médecine : 22 143 documents
— Droit : 9 964 documents

Pour le besoin de notre expérimentation et pour permettre une évaluation par certaines métriques, il est nécessaire 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 définit trois groupes de mots, comme le montre le tableau 5.4 : 1) le groupe des mots communs, ceux appartenant à 3 catégories ou plus (ils représentent les 12 premiers pour cent des mots du corpus triés par occurrence) ; 2) les groupes de mots relevant de deux catégories ; 3) ainsi que ceux appartenant à une catégorie unique.

	HISTOIRE	ECONOMIE	INFORMATIQUE	MEDECINE	DROIT
HISTOIRE	394 286	49 387	16 007	35 752	14 523
ECONOMIE		80 868	12 664	5 204	3 669
INFORMATIQUE			60 614	2 196	931
MEDECINE				74 859	1 209
DROIT					14 920

Total des mots rencontrés dans 1 seul thème : 625 547
Total des mots rencontrés dans 2 thèmes : 141 542
Total des mots rencontrés dans 3 thèmes et + : 110 441

Tableau 5.4. *Présentation du corpus extrait de Wikpedia. Les mots rencontrés dans 3 thèmes, aussi appelés "mots communs", représentent environ 12% des mots du corpus triés par occurence.*

La figure 5.5 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 de mots sont respectées lors de l'insertion, et seuls les mots outils sont supprimés.

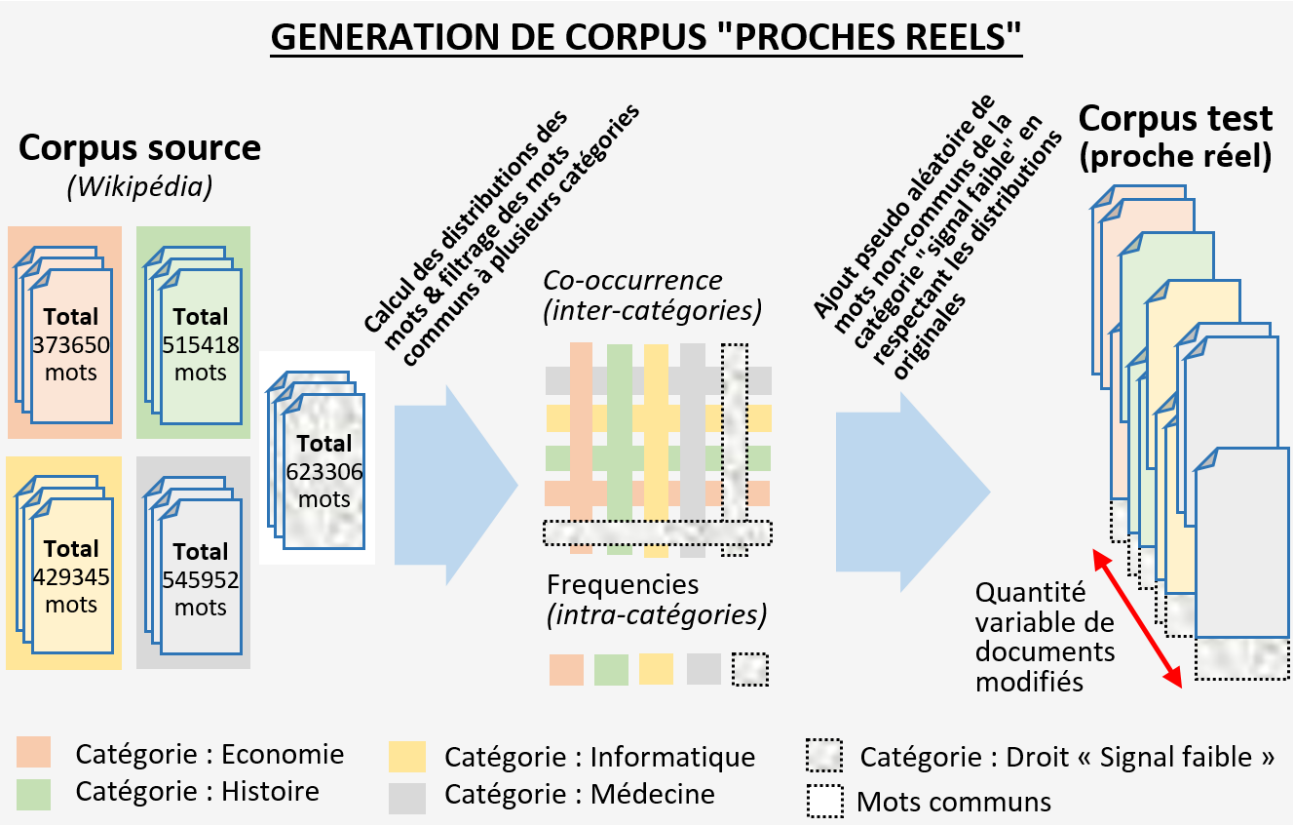

FIGURE 5.5. *Méthode de génération du corpus "proche du réel" qui consiste en l'injection de mots dits "non-communs" empruntés au thème "signal faible" (Droit en l'occurence) en respectant la distribution originale. Ces mots non-communs sont identifiés par une étude de co-occurrence entre thèmes.*

Pour ce test, nous avons utilisé un modèle *Word2Vec* pré-entrainé sur le corpus français de Wikipédia (Dump du 07/11/2016) (Schöch, 2016). L'identification du cluster *signal faible* s'effectue de la façon suivante :

— Au préalable, l'appartenance de chaque mot aux 5 thèmes a été effectué à partir du corpus (e.g. le mot "donnée" appartient aux thèmes Histoire, Médecine et Informatique, le mot "avocat" appartient au thème Droit) ;
— Pour chaque cluster obtenu, nous calculons la somme des poids des mots n'appartenant qu'à un seul thème (e.g. le mot "donnée" ne sera pas pris en compte) ; Pour chaque cluster, nous obtenons une valeur pour chacun des 5 thèmes ;
— Le *signal faible* (Droit) est détecté lorsque la valeur du thème Droit est la plus importante.

Toutes les données (documents de Wikipédia français et modèle pré-entrainé *Word2Vec*) utilisées dans ce travail sont accessibles en suivant ces liens : `https://doi.org/10.5281/zenodo.3260045` (Maitre, 2019a), `http://doi.org/10.5281/zenodo.162792` (Schöch, 2016)

Chaque jeu de données se compose de 250 documents de chaque thème. Nous insérons dans un nombre variable de documents 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. Il est choisi empiriquement après plusieurs expérimentations (échantillonnage de paramètres) sur un sous-ensemble de données. Dans nos expériences, les impacts de la variation de cette valeur ne sont pas très sensibles lorsqu'elle appartient à l'intervalle [0.6-0.9]. Il permet d'obtenir les meilleures valeurs de cohérence pour les clusters ainsi que la meilleure détection des clusters de *signaux faibles*. Cette valeur affecte cependant le nombre de clusters détectés. Les mots du *signal faible* sont insérés en respectant les distributions précédemment calculées sur le corpus de documents. Le nombre de documents avec le *signal faible* varie de 100 à 800 par pas de 50 documents. Nous effectuons ce test 10 fois.

Les résultats obtenus (cf. Figure 5.6), montrent la robustesse de l'algorithme 2 même pour un très petit nombre de mots injectés de la catégorie "Droit" (*signal faible*) comparé à chaque *LDA* pour $k \in \{2 \dots 8\}$. 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 (12 mots par document). 0.82% correspond à 3 600 mots (12 mots injectés dans 300 documents). Chaque fois que nous trouvons le cluster *signal faible*, nous recherchons celui qui a la valeur de cohérence la plus élevée. Dans la figure 5.7, nous montrons que l'algorithme 2 peut détecter le cluster *signal faible* (i.e, celui détecté comme *signal faible* est celui qui présente le plus de cohérence pour tous les niveaux de *LDA* $k \in \{2 \dots 8\}$). L'algorithme *LDA* seul donne parfois une partition où le cluster *signal faible* est présent (avec une valeur de cohérence de similarité inférieure, cf. Figure 5.7). Ce test montre donc l'intérêt et la contribution de cette étude dans la détection d'un *signal faible* par une approche conjointe *LDA/Word2Vec*.

L'insertion de mots au sein des documents modifie cependant leurs contenus. Le corpus tend à simuler ce que pourrait être des documents dans lesquelles quelques phrases avec des mots spécifiques d'un *signal faible* (i.e. patterns) sont présents.

Nous donnerons dans la section suivante, une autre approche reposant sur la valeur de cohérence et la pondération *tf-idf* pour déterminer le *signal faible* parmi les clusters trouvés.

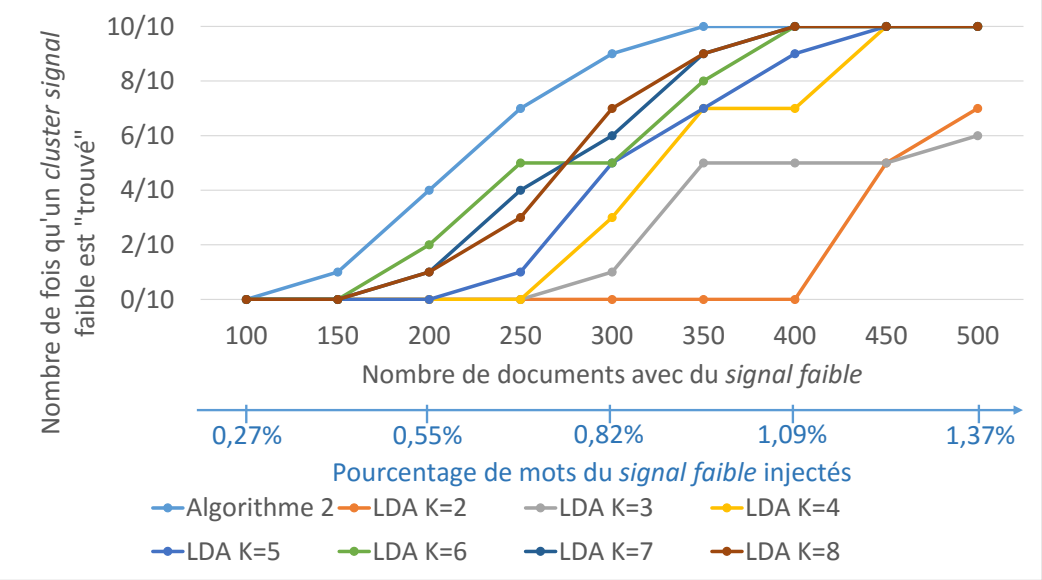

FIGURE *5.6.* *Résultats de l'algorithme 2 comparé aux 7 LDAs originaux paramétrés avec k variant de 2 à 8 sur la détection d'un cluster signal faible dans les clusters déterminés par l'approche conjointe. Dans chaque document, nous insérons 3 séries de 4 mots de la catégorie "Droit" (signal faible).*

5.2.2 Test sur des documents médicaux

Nous proposons dans cette section d'étudier l'efficacité de notre solution sur le corpus Ohsumed [5] pressenti pour être plus spécifique. Ce dernier est composé de 56 984 comptes rendus médicaux, chacun relatif à une pathologie parmi 23 présentes dans le corpus (Hersh *et al.*, 1994).

Nous proposons d'ajouter artificiellement une 24ème pathologie/catégorie non déjà représentée dans le corpus. Elle fera office de *signal faible*. Pour cela, nous choisissons des documents de référence issus de Wikipédia [6] traitant de la maladie Ebola et de maladies connexes. Le jeu de données supplémentaire est accessible en suivant ce lien : `https://doi.org/10.5281/zenodo.3591580` (Maitre, 2019b)

Nous injectons alors, sur un volume total du corpus de 70 Mo, 500 Ko de documents portant sur ces agents infectieux déclencheurs d'épidémie. La méthode a été évaluée en faisant varier k sur un intervalle allant de 15 à 35. L'ensemble du Wikipédia Anglais a été utilisé pour entraîner un modèle de type *Word2Vec* à l'aide de l'outil

5. Disponible sur : `http://disi.unitn.it/moschitti/corpora.htm`
6. Wikipedia anglais, dump du 26/11/2018.

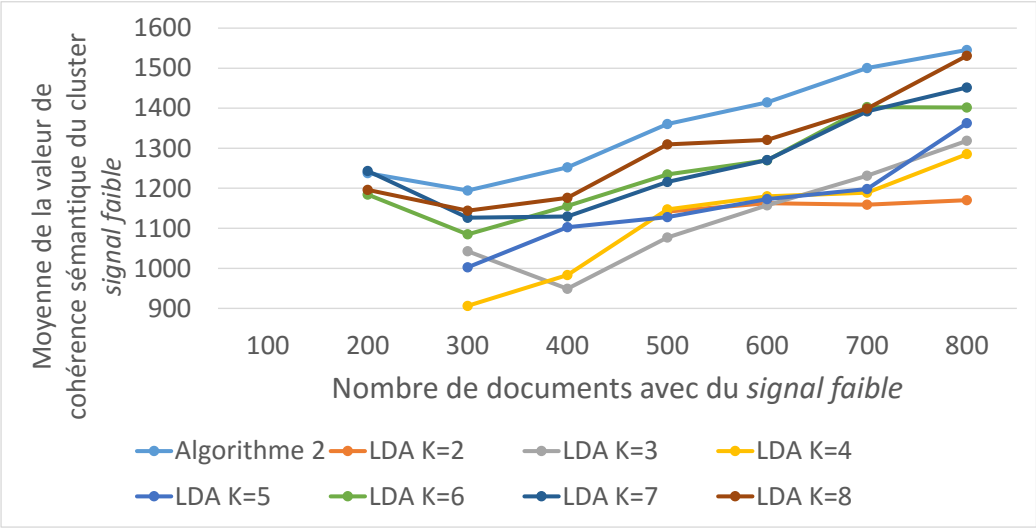

FIGURE 5.7. *Résultats obtenus à partir de 10 tests de l'algorithme 2 comparés aux 7 LDAs originaux paramétrés avec k variant de 2 à 8. L'algorithme 2 détecte le cluster signal faible avec la plus grande valeur de cohérence comparée à celles de LDA.*

Word2vec-on-wikipedia [7]. Nos tests utilisent la méthode de pondération "*tf-idf*" pour choisir les mots sur lesquels calculer la cohérence des clusters et construrie l'arborescence des clusters (à la place de la pondération de *LDA*) (cf. Equation 5).

$$\textit{tf-idf}_i = f_{t_i,\{d_j:t_i\in d_j\}} \cdot \log \frac{|D|}{\{d_j : t_i \in d_j\}} \quad \text{où} \quad f_{t_i,\{d_j:t_i\in d_j\}} = \frac{n_i}{\{d_j : t_i \in d_j\}} \qquad [5]$$

tf-idf$_i$ représente la valeur de pondération du mot t_i dans le corpus de documents D, $|D|$ le nombre total de documents dans le corpus, $\{d_j : t_i \in d_j\}$ le nombre de documents où le terme t_i apparaît et n_i le nombre d'occurrence de t_i dans D.

Une fois cette pondération appliquée sur l'ensemble des mots de chaque cluster, ceux-ci sont triés par ordre décroissant de cette valeur. La cohérence de chaque cluster est alors calculée sur les 10 premiers mots. Les valeurs de k pour *LDA* sont [15,20,25,27,30,32,35]. Le seuil pour l'élagage de l'arborescence est arbitrairement fixé à 0.40. L'arbre construit possède donc 7 niveaux et 184 clusters.

Nom du cluster	c_1 / 13	c_2 / 13	c_3 / 13	c_4 / 13	c_5 / 13	...
Cohérence du cluster (I1)	32.59	29.94	25.29	21.03	19.94	...
LDA k =	15	25	20	20	20	...
Premiers mots du cluster	scurfy pcnsl spm dracunculiasis coc aerd agep alkaloidal agn rosacea	paroxetine ntds gbp dexmedetomidine mor mefenamic deslorelin alfuzosin btb dabao	ebola chikungunya song deworming announces soil-transmitted vinson ebolavirus ntd ntds	mvb pni litho qsp pvri rsi vge duncan nmd impactor	ngc oxy efamol cmh fet nrv tpcs phr parvalbumin mcj	...

Tableau 5.5. *Expérimentation sur un corpus de compte-rendu médicaux avec ajout d'un signal faible sous forme de documents portant sur des agents infectieux. Présentation des résultats obtenus sur les 5 premiers clusters (parmi les 13 clusters) obtenus après construction de l'arbre et élagage. Pour chaque cluster, est indiqué la valeur de cohérence selon l'indicateur I_1, la valeur de k du LDA où il est détecté ainsi que les 10 premiers mots triés par tf-idf.*

7. Disponible sur : https://github.com/jind11/word2vec-on-wikipedia

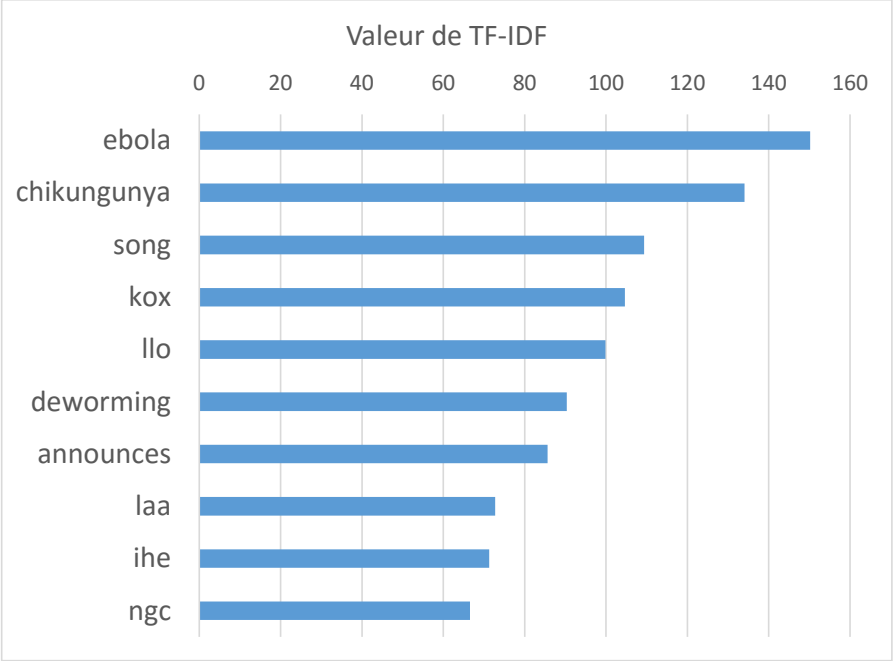

FIGURE 5.8. *Liste des mots du corpus ayant les grandes valeurs de pondération tf-idf. On remarque que les mots clés du cluster signal faible détecté sont parmi les premiers de la liste.*

Le tableau 5.5 montre les résultats obtenus sur le corpus Ohsumed modifié. Après élagage, 13 clusters sont retenus. On remarque que les mots relatifs aux documents portant sur les maladies infectieuses tropicales (*signal faible*) sont capturés par un cluster classé 3ème / 13 en termes de cohérence. Les mots courts semblent inintelligibles mais il s'agit d'acronymes médicaux. Le fait que le cluster *signal faible* ne possède pas la cohérence la plus élevée (mais seulement la 3ème sur 13) parmi les clusters détectés de l'arbre (après élagage) provient de la nature même des documents injectés. Elles décrivent en effet un spectre large de maladies infectieuses contrairement aux autres comptes-rendus qui s'attachent à une description centrée sur une pathologie.

La figure 5.8 montre les 10 premiers mots de l'ensemble du corpus triés selon leur valeur de pondération *tf-idf*. Parmis les 10 premiers mots, 5 appartiennent au thème *signal faible*. Les résultats montrent la pertinence de notre solution (indicateurs et algorithme) dans la recherche de *signaux faibles* dans ce contexte de corpus médical augmenté de documents ayant trait à des maladies infectieuses tropicales.

Parmis les 10 premiers mots du corpus, 5 appartiennent au cluster *signal faible*. L'exploitation combinée des clusters résultats de l'algorithme 2 avec la liste des mots du corpus triés par leur valeur de pondération *tf-idf* permet la détection du cluster *signal faible*. Les mots du *signal faible* connus, il est alors possible d'enrichir cette liste par des phases d'exploration sur le réseau.

5.3 Synthèse des tests

Pour l'analyse des documents et l'extraction du *signal faible*, nous adoptons une approche conjointe *LDA*/*Word2Vec*. 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 grâce à un critère de cohérence calculé à partir de *Word2Vec* 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*. Pour détecter *in fine* ce dernier, le même critère de cohérence est utilisé mais uniquement sur les mots rares des clusters.

Les différents tests montrent que l'approche conjointe conduit à la sélection des clusters cohérents au sens de la définition du *signal faible* ainsi qu'à la détection du cluster porteur du *signal faible* lorsque ce critère est appliqué uniquement sur les mots rares (mots non-communs). L'approche *LDA* conditionne la recherche de clusters situés sur le même niveau d'arborescence (k fixé *a priori*) : certains clusters peuvent être cohérents et d'autres non à ce niveau de décomposition. La recherche des clusters les plus pertinents nécessite une analyse approfondie de l'arbre de partitionnement, ce qui est effectuée par l'algorithme 2.

L'approche conjointe *LDA/Word2Vec* répond aux caractéristiques retenus du *signal faible* :

— le *signal faible* est caractérisé par un faible nombre de mots par document et présents dans peu de documents (rareté, anormalité) : la méthode d'injection des mots-clés du *signal faible* dans les documents du corpus pour la réalisation des tests est conforme à cette prérogative. De plus, le critère de détection du *signal faible* est appliqué uniquement sur les mots rares.

— le *signal faible* 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) : l'algorithme repose sur un critère de cohérence construit pour mettre en évidence les propriétés contextuelles des mots clés du *signal faible*, et ainsi capturer dans un cluster les associations très locales. L'élagage permet d'isoler des autres clusters ceux susceptibles de contenir le *signal faible* (nous ne présupposons pas que les thèmes puissent être décrits d'une manière hiérarchique i.e. nous n'utilisons pas par exemple *hLDA*).

L'étape de visualisation, qui vient après l'étape de clustering multi-niveaux dans la chaine de traitement, permet de mettre en évidence les documents 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 construites 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 5.9 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 5.10 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.

6 Conclusion

Cette étude porte sur la recherche des mots-clés relevant d'un thème *signal faible* éventuellement présent dans un corpus de documents transmis par un lanceur d'alerte. L'étude décrite dans l'article porte sur la première phase de la chaine de traitements de la plateforme, autrement dit, l'analyse des documents reçus doit simultanément permettre de : (1) découvrir les thèmes ou l'arborescence des thèmes portés par les documents ; (2) classer les documents relativement aux thèmes ; (3) détecter les mots-clés pertinents composants les thèmes, (4) et enfin, découvrir les mots-clés du thème *signal faible*.

Nous précisons la définition du *signal faible* et ses caractéristiques. Elle est essentielle en effet, car elle conditionne le critère d'évaluation qui permet de déterminer quel cluster final a capturé les mots clés relevant du *signal faible*. Comme nous le montrons dans l'article, elle justifie l'approche conjointe modèle thématique / plongement lexical, et contraint le choix des méthodes utilisées : *LDA* et *Word2Vec*. Cette dernière permet notamment de construire un critère permettant de capturer les régularités sémantiques et ainsi de mettre en évidence les mots clés apparaissant dans des contextes similaires. Il semble donc adapter pour identifier les mots-clés du *signal faible*, contrairement aux critères globaux utilisés seuls qui ne prennent pas en compte le contexte des mots.

Pour détecter si un cluster relève du *signal faible*, il est nécessaire que son critère de cohérence soit le plus élevé, relativement aux autres clusters retenus. Dans la première expérience sur les données artificielles, on s'attache

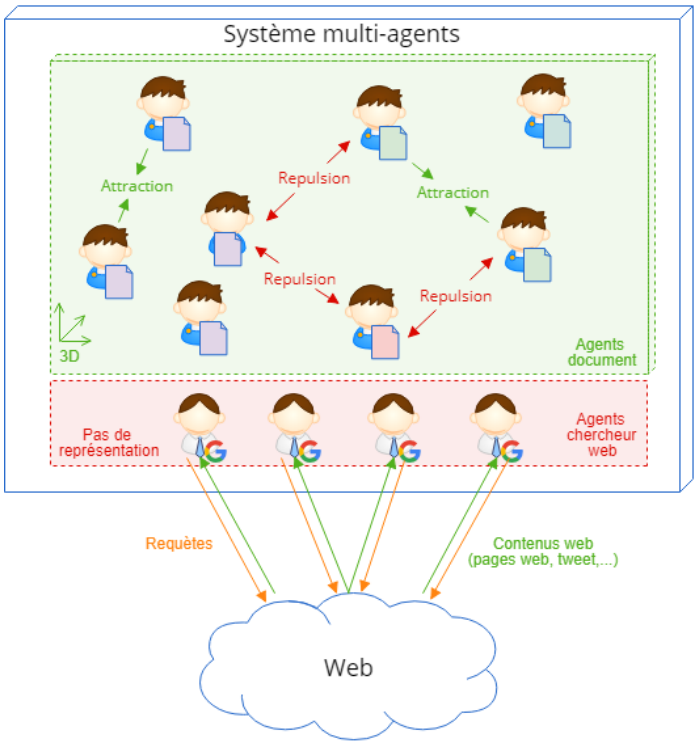

FIGURE *5.9.* *Les agents document interagissent les uns avec les autres grâce à des actions de type attraction/répulsion. Les agents de recherche construisent des requêtes à partir des mots associés au signal faible.*

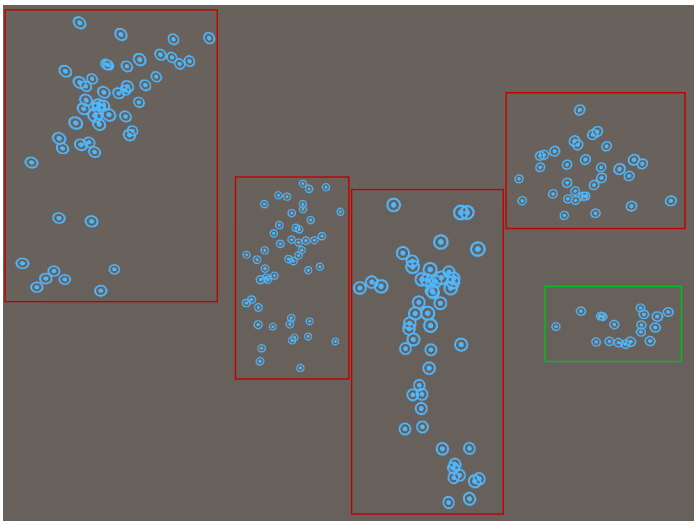

FIGURE *5.10.* *Cette figure représente la capacité du système à organiser les documents dans un espace 3D et à faire émerger des clusters. La boîte rouge représente des clusters de documents relatifs aux thèmes principaux et la boîte verte représente ceux relatifs au signal faible.*

principalement à montrer la pertinence du critère contextuel construit sur *Word2Vec* L'objectif du jeu de données artificiel est de réaliser une preuve de concept pour évaluer l'approche proposée dans l'article. Ce jeu synthétique permet de vérifier, dans un cadre contrôlé, notre approche et que les résultats soient conformes à ceux attendus.

Nous réalisons des tests proches du réel avec le corpus de données Wikipédia. Celui-ci est composé de documents appartenant à 5 catégories de la version française de Wikipédia. Nous respectons, lors de l'injection du *signal faible*, les fréquences des mots présentent dans les documents de Droit. Ainsi on s'assure que notre corpus respecte les mêmes caractéristiques fréquentielles que les documents Wikipédia. Les résultats ont montré que notre algorithme est capable de retrouver le *signal faible* même avec un faible nombre de mots injectés du thème.

Pour s'approcher cependant d'un jeu plus réaliste, nous proposons une troisième expérimentation sur une base de données de comptes rendus médicaux dans laquelle nous injectons des documents de même type mais portant sur des agents infectieux déclencheurs d'épidémie (e.g. virus Ebola). Les caractéristiques du jeu de données sont présentées dans l'article ainsi que les résultats. La dépendance entre mots-clés d'un *signal faible* est une des caractéristiques essentielles pour discriminer ces mots-clés. C'est ce que le critère contextuel construit sur *Word2Vec* cherche à mettre en évidence. Les travaux futurs porteront sur le developpement d'un module additionnel cherchant à mettre en corrélation les informations portées par un *signal rare* avec un contexte informationnel plus large grâce à des phases d'exploration sur les réseaux.

Références

AH-PINE J., LEMOINE J. & BENHADDA H. (2005). Un nouvel outil de classification non supervisée de documents pour la découverte de connaissances et la détection de signaux faibles : RARES Text^{TM}. In *Journées sur les systèmes d'information élaborée*, Ile Rousse, France.

ALGHAMDI R. & ALFALQI K. (2015). A Survey of Topic Modeling in Text Mining. *IJACSA) International Journal of Advanced Computer Science and Applications*, 6(1), 147–153.

ANSOFF H. I. (1975). Managing Strategic Surprise by Response to Weak Signals. *California Management Review*, 18(2), 21–33.

BAKAROV A. (2018). A Survey of Word Embeddings Evaluation Methods.

BAKAROV A. & GUREENKOVA O. (2018). Automated detection of non-relevant posts on the russian imageboard "2ch" : Importance of the choice of word representations. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 10716 LNCS, p. 16–21 : Springer Verlag.

BANSAL M., GIMPEL K. & LIVESCU K. (2014). Tailoring continuous word representations for dependency parsing. In *52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014 - Proceedings of the Conference*, volume 2, p. 809–815.

BAUMEL T., COHEN R. & ELHADAD M. (2016). Sentence Embedding Evaluation Using Pyramid Annotation. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, p. 145–149.

BHATTACHARYYA A. (1943). On A Measure of Divergence Between Two Statistical Populations Defined by their Probability Distributions. *Bulletin of the Calcutta Methematical Society*, 35(1), 99–109.

BLEI D. M., GRIFFITHS T. L., JORDAN M. I. & TENENBAUM J. B. (2004). Hierarchical topic models and the nested Chinese restaurant process. In *Advances in Neural Information Processing Systems* : Neural information processing systems foundation.

BLEI D. M. & LAFFERTY J. D. (2006). Dynamic topic models. In *ACM International Conference Proceeding Series*, volume 148, p. 113–120.

BLEI D. M., NG A. Y. & JORDAN M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3(4-5), 993–1022.

CLIFTON L., CLIFTON D. A., WATKINSON P. J. & TARASSENKO L. (2011). Identification of patient deterioration in vital-sign data using one-class support vector machines. In *2011 Federated Conference on Computer Science and Information Systems, FedCSIS 2011*, p. 125–131.

COFFMAN B. (1997). Weak signal research, part I : Introduction. *Journal of Transition Management*, 2(1).

COLLOBERT R., WESTON J., BOTTOU L., KARLEN M., KAVUKCUOGLU K. & KUKSA P. (2011). Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12, 2493–2537.

DAY G. S. & SCHOEMAKER P. J. H. (2005). Scanning the Periphery. *Harvard Business Review*, 83(11), 135–148.

DECKER R., WAGNER R. & SCHOLZ S. W. (2005). An internet-based approach to environmental scanning in marketing planning. *Marketing Intelligence & Planning*, 23(2), 189–199.

DEERWESTER S., DUMAIS S. T. & HARSHMAN R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6), 391–407.

DOMINGUES R., FILIPPONE M., MICHIARDI P. & ZOUAOUI J. (2018). A comparative evaluation of outlier detection algorithms : Experiments and analyses. *Pattern Recognition*, 74, 406–421.

EBRAHIMKHANLOU A. & SALAMONE S. (2017). A probabilistic framework for single-sensor acoustic emission source localization in thin metallic plates. *Smart Materials and Structures*, 26(9).

GLOBERSON A., CHECHIK G., PEREIRA F. & TISHBY P. N. (2007). Euclidean Embedding of Co-occurrence Data. *Journal of Machine Learning Research*, 8, 2265–2295.

GUNES V., MENARD M. & PETITRENAUD S. (2010). Multiple classifier systems : tools and methods. In C. CHEN, Ed., *Handbook of Pattern Recognition and Computer Vision*, p. 23–46. World Scientific.

HERSH W., BUCKLEY C., LEONE T. J. & HICKAM D. (1994). OHSUMED : An interactive retrieval evaluation and new large test collection for research. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 1994*, p. 192–201 : Association for Computing Machinery, Inc.

HILTUNEN E. (2008). The future sign and its three dimensions. *Futures*, 40(3), 247–260.

HOFMANN T. (2001). Unsupervised learning by probabilistic Latent Semantic Analysis. *Machine Learning*, 42(1-2), 177–196.

KAKKONEN T., MYLLER N., SUTINEN E. & TIMONEN J. (2008). Comparison of dimension reduction methods for automated essay grading. *Educational Technology and Society*, 11(3), 275–288.

KIM J. & LEE C. (2017). Novelty-focused weak signal detection in futuristic data : Assessing the rarity and paradigm unrelatedness of signals. *Technological Forecasting and Social Change*, 120(June 2016), 59–76.

KÖHN A. (2016). Evaluating Embeddings using Syntax-based Classification Tasks as a Proxy for Parser Performance. p. 67–71 : Association for Computational Linguistics (ACL).

LEVY O. & GOLDBERG Y. (2014a). Linguistic regularities in sparse and explicit word representations. In *CoNLL 2014 - 18th Conference on Computational Natural Language Learning, Proceedings*, p. 171–180.

LEVY O. & GOLDBERG Y. (2014b). Neural word embedding as implicit matrix factorization. In *Advances in Neural Information Processing Systems*, volume 3, p. 2177–2185 : Neural information processing systems foundation.

MAITRE J. (2019a). A Wikipedia dataset of 5 categories.

MAITRE J. (2019b). A Wikipedia dataset of Ebola disease related articles.

MÉNARD M. (2001). Fuzzy clustering and switching regression models using ambiguity and distance rejects. *Fuzzy Sets and Systems*, 122(3), 363–399.

MÉNARD M. & EBOUEYA M. (2002). Extreme physical information and objective function in fuzzy clustering. *Fuzzy Sets and Systems*, 128(3), 285–303.

MIKOLOV T., CORRADO G., CHEN K. & DEAN J. (2013a). Efficient Estimation of Word Representations in Vector Space. In *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings* : International Conference on Learning Representations, ICLR.

MIKOLOV T., SUTSKEVER I., CHEN K., CORRADO G. & DEAN J. (2013b). Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems* : Neural information processing systems foundation.

MOHAMMADI-GHAZI R., MARZOUK Y. M. & BÜYÜKÖZTÜRK O. (2018). Conditional classifiers and boosted conditional Gaussian mixture model for novelty detection. *Pattern Recognition*, 81, 601–614.

NALLAPATI R. M., AHMED A., XING E. P. & COHEN W. W. (2008). Joint latent topic models for text and citations. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, p. 542–550.

PARK C. & CHO S. (2017). Future sign detection in smart grids through text mining. *Energy Procedia*, 128, 79–85.

RAMEZANI R., ANGELOV P. & ZHOU X. (2008). A fast approach to novelty detection in video streams using recursive density estimation. In *2008 4th International IEEE Conference Intelligent Systems, IS 2008*, volume 3, p. 142–147.

RIGOUSTE L., CAPPÉ O. & YVON F. (2006). Quelques observations sur le modèle LDA. *Journées internationales d'Analyse statistique des Données Textuelles*, 8, 819–830.

SCHNABEL T., LABUTOV I., MIMNO D. & JOACHIMS T. (2015). Evaluation methods for unsupervised word embeddings. In *Conference Proceedings - EMNLP 2015 : Conference on Empirical Methods in Natural Language Processing*, p. 298–307.

SCHÖCH C. (2016). A word2vec model file built from the French Wikipedia XML Dump using gensim.

SHEN Z. Y., SUN J. & SHEN Y. D. (2008). Collective latent dirichlet allocation. In *Proceedings - IEEE International Conference on Data Mining, ICDM*, p. 1019–1024.

SOCHER R., BAUER J., MANNING C. D. & NG A. Y. (2013a). Parsing with compositional vector grammars. In *ACL 2013 - 51st Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, volume 1, p. 455–465 : Association for Computational Linguistics (ACL).

SOCHER R., PERELYGIN A., WU J. Y., CHUANG J., MANNING C. D., NG A. Y. & POTTS C. (2013b). Recursive deep models for semantic compositionality over a sentiment treebank. In *EMNLP 2013 - 2013 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, p. 1631–1642 : Association for Computational Linguistics (ACL).

THORLEUCHTER D. & VAN DEN POEL D. (2013). Weak signal identification with semantic web mining. *Expert Systems with Applications*, 40(12), 4978–4985.

TSVETKOV Y., BOYTSOV L., GERSHMAN A., NYBERG E. & DYER C. (2014). Metaphor detection with cross-lingual model transfer. In *52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014 - Proceedings of the Conference*, volume 1, p. 248–258.

TSVETKOV Y., FARUQUI M., LING W., LAMPLE G. & DYER C. (2015). Evaluation of word vector representations by subspace alignment. In *Conference Proceedings - EMNLP 2015 : Conference on Empirical Methods in Natural Language Processing*, p. 2049–2054.

TURIAN J., RATINOV L. & BENGIO Y. (2010). Word representations : A simple and general method for semi-supervised learning. In *ACL 2010 - 48th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, p. 384–394.

VAN DER MAATEN L. & HINTON G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9, 2579–2625.

YOON J. (2012). Detecting weak signals for long-term business opportunities using text mining of Web news. *Expert Systems with Applications*, 39(16), 12543–12550.

ZHAO W., CHEN J. J., PERKINS R., LIU Z., GE W., DING Y. & ZOU W. (2015). A heuristic approach to determine an appropriate number of topics in topic modeling. *BMC Bioinformatics*, 16(13).

