Détection et analyse des signaux faibles. Développement d'un framework d'investigation numérique pour un service caché Lanceurs d'alerte

Julien Maitre

LA ROCHELLE UNIVERSITÉ

ÉCOLE DOCTORALE EUCLIDE

Laboratoire L3i (Informatique, Image et Interaction)

THÈSE présentée par :

Julien MAITRE

soutenue le : **6 avril 2022**
pour obtenir le grade de : **Docteur de La Rochelle Université**
Discipline : **Informatique et Applications**

Détection et analyse des signaux faibles. Développement d'un framework d'investigation numérique pour un service caché Lanceurs d'alerte.

JURY :

Nicole VINCENT	Professeure, Université Paris Descartes, Rapportrice
Florence SEDES	Professeure, Université de Toulouse, Rapportrice
Guillaume CHIRON	Ingénieur de recherche, Entreprise ARIADNEXT, Examinateur
Ronan CHAMPAGNAT	Maître de conférences, La Rochelle Université, Examinateur
Michel MÉNARD	Professeur, La Rochelle Université, Directeur de thèse
Alain BOUJU	Maître de conférences, La Rochelle Université, Coencadrant de thèse

Remerciements

Tout d'abord, je tiens à remercier mes encadrants, **Michel MÉNARD** (Professeur à La Rochelle Université), **Alain BOUJU** (Maître de conférences à La Rochelle Université) et **Guillaume CHIRON** (Ingénieur de recherche à l'entreprise ARIADNEXT) pour avoir proposé ce sujet, avec lesquels j'ai pris un grand plaisir à travailler et sans qui rien n'aurait été possible. J'espère que nous aurons l'occasion de collaborer à nouveau ensemble.

Je remercie les membres du jury d'avoir accepté d'évaluer ces travaux de thèse :

— **Florence SEDES** (Professeure à l'Université de Toulouse) qui fut la présidente du jury et rapportrice ;
— **Nicole VINCENT** (Professeure à l'Université Paris Descartes) pour son travail en tant que rapportrice ;
— **Ronan CHAMPAGNAT** (Maître de conférences à La Rochelle Université) et **Guillaume CHIRON** pour leurs rôle d'éxaminateurs ;
— mes encadrants, **Michel MÉNARD**, **Alain BOUJU** qui ont également pris part à ce jury.

Je me dois également de remercier **Yacine GHAMRI-DOUDANE** (Directeur du laboratoire), et l'ensemble des chercheurs rattachés au L3i pour m'avoir offert tout le confort, le soutien matériel nécessaire, mais surtout pour l'environnement de travail agréable qui permet aux doctorants et futurs docteurs, de travailler dans de bonnes conditions.

Je tiens également à remercier mes financeurs, sans lesquels cette thèse n'aurait pas été possible : l'ex-région Poitou-Charentes, intégrée dans la Nouvelle Aquitaine.

Merci à toutes les personnes qui m'ont aidé durant cette thèse, que ce soit sur l'aspect technique ou lors de nos échanges.

Plus personnellement, je tiens à remercier toutes les personnes qui ont fait de ce doctorat une période enrichissante et épanouissante. Ces moments étaient une bouffée d'air frais qui m'ont aidé à me changer les idées le temps d'un week-end, d'une soirée ou pour quelques instants.

Un remerciement particulier à mes collègues de bureau, anciens et nouveaux, qui m'ont supporté durant toutes ces années : Van, Iuliia, Zuheng, Khoa, Yasmine, Nam, Imane.

Je conclus mes remerciements par mes parents pour leur soutien indéfectible et leur aide pour les nombreuses relectures.

Table des matières

Remerciements **i**

Introduction générale **5**

Contexte général et positionnement 5

Sélection des *signaux faibles* 8

Proposition d'une définition d'un *signal faible* 11

Contributions 12

1 Définitions et Positionnement Scientifique **19**

1.1 Introduction 19

1.2 Angles d'analyse utilisés 20

1.2.1 Analyse selon les objectifs de détection 21

1.2.2 Analyse selon la source de données 28

1.2.3 Analyse selon les domaines d'analyse 29

1.2.4 Techniques de fouille de données/extraction de connaissances 33

1.2.5 Evaluation des techniques 55

1.3 Positionnement 55

2 Modélisation thématique, plongement de mots et exploration d'une collection de documents **61**

2.1 Introduction 62

2.2 Modèle thématique 63

2.2.1 L'analyse sémantique latente 63

2.2.2 L'analyse sémantique latente probabiliste 64

2.2.3 L'allocation de Dirichlet latente . . . 65
2.3 Word embedding . . . 66
2.4 Justification de l'approche *LDA*... . . . 68
2.5 Justification d'une approche conjointe... . . . 70
2.6 *LDA* augmenté avec *Word2Vec* . . . 71
2.6.1 Cas d'utilisation de *LDA* sur Wikipédia . . . 73
2.6.2 Indicateur de cohérence en tant que mesure intra-thème . . . 75
2.6.3 Recherche du paramètre k conduisant aux thèmes les plus pertinents au sens du critère de cohérence . . . 76
2.6.4 Une approche heuristique pour déterminer les thèmes les plus pertinents sur l'ensemble de l'arborescence *LDA* . . . 77
2.7 Expérimentations . . . 79
2.7.1 Tests sur un corpus artificiel . . . 80
2.7.2 Tests sur des corpus de données réelles . . . 89
2.8 Conclusion . . . 97

3 *Agent mining* et développement d'un logiciel 99

3.1 Introduction . . . 100
3.2 Etat de l'art . . . 101
3.2.1 Extraction de connaissances . . . 101
3.2.2 Système Multi-Agents . . . 107
3.2.3 *Agent mining* : *Data mining* et Système multi-agents . . . 115
3.3 Système multi-agents et *data mining* proposé . . . 115
3.3.1 Chaine de traitement . . . 116
3.3.2 Système multi-agents associé aux documents . . . 116
3.3.3 Système multi-agents de recherche . . . 123
3.3.4 Système multi-agents associé aux mots . . . 127
3.3.5 Analyse de l'évolution des thèmes . . . 128
3.4 Présentation de l'architecture du logiciel WILD . . . 129
3.4.1 Composants et services du logiciel WILD . . . 131

3.4.2 Composants et services du système d'investigation 134
3.5 Critères et paramétrage de la chaine de traitement 138
3.6 Conclusion 142

4 Expérimentations 149
4.1 Introduction 149
4.2 Projets H2020 [2014-2021] 150
4.2.1 Corpus 150
4.2.2 Résultats 151
4.2.3 Conclusion 162
4.3 Analyse bibliographique d'une base de données d'articles 164
4.3.1 Corpus 164
4.3.2 Mise en œuvre 165
4.3.3 Principaux résultats obtenus 167

Conclusion générale et perspectives 169

Annexe A Critères et paramètrage des expérimentations 175

Annexe B Documents identifiés dans les domaines de la santé et du médical 179

Annexe C Paradigmes organisationnels dans les systèmes multi-agents 185

Annexe D Expérimentation complémentaire projets H2020 187

Publications 193

Table des Figures 195

Liste des Tableaux 205

Liste des Algorithmes 209

Bibliographie 211

Introduction générale

Contexte général et positionnement

La place du numérique dans notre société moderne engendre des mutations importantes, rapides et massives qui touchent tous les secteurs d'activités (commerciaux, industriels, économiques, éducatifs, de la santé, de la culture, du transport, de l'énergie...). Ces mutations modifient les activités humaines dans leurs usages (production de contenus, interactions homme machines, ...), les équipements (smartphones, tablettes, ...) et les organisations (réseaux sociaux, communautés virtuelles, ...). Cette évolution soulève de nombreuses questions technologiques et sociétales : enjeux technologiques, managériaux, juridiques, environnementaux.

Cette explosion du numérique suscite la nécessité de mise en œuvre de nouvelles interactions avec l'environnement et suggère le développement de nouvelles orientations scientifiques : acquisition de données, Interface Homme-Machine (IHM), problématique de traitement de données réparties, ...

Des problématiques de structuration de contenus dynamiques, complexes, hétérogènes, multiples nécessitent des solutions intégrant une contextualisation des traitements, l'utilisation de représentations sémantiques, le développement de méthodes de raisonnements adaptatifs et d'apprentissage automatique, et le déploiement de nouvelles modalités d'interactions Homme-Machine.

L'objectif principal de ce travail est d'aider à la prise de décision face à l'augmentation drastique des signaux transmis pour les décideurs 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. Pour les entreprises, ces dynamiques représentent pourtant des opportunités de croissance [VBV10] ou bien des menaces fondamentales [vdGVD10] sur lesquelles elles doivent bâtir leur orientation stratégique, facteur clé, face au défi permanent de maintenir leur avantage concurrentiel. Elles offrent cependant des opportunités pour

comprendre les évolutions politiques, économiques, sociales et technologiques. Il est donc nécessaire de développer, sous la forme de plateformes d'investigation (comme par exemple, dans le contexte des lanceurs d'alertes, la plateforme de la 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 algorithmes robustes pour détecter les *signaux faibles*, extraire, analyser les informations fournies par ces derniers et s'ouvrir à un contexte d'information plus large.

Du point de vue applicatif, nous inscrivons cette étude dans le cadre du data journalisme, dont les acteurs 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 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 [1], 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 [2] au départ invisible. Dans notre étude, pour un problème donné, nous faisons l'hypothèse qu'il existe des patterns caractéristiques de *signaux faibles* (non encore identifiés à la reception des premiers documents précusseurs des *signaux faibles*) qui permettent de détecter des corrélations non visibles entre documents.

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 à partir d'autres sources d'information.
- **A3** : Visualisation analytique.

1. `https://wikileaks.org/`
2. `https://wardiaries.wikileaks.org/`

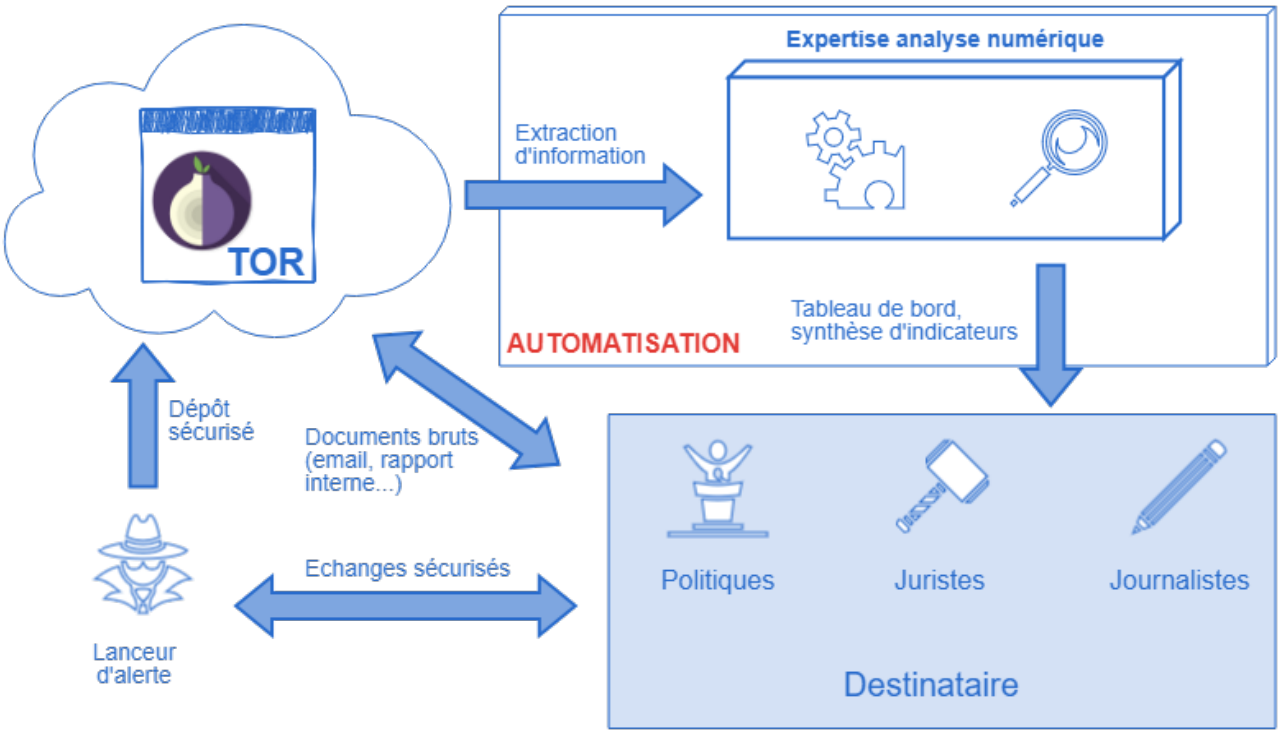


FIGURE *1* – ***Aperçu d'un exemple de plateforme d'investigation. Dans le contexte du data journalisme, elle doit répondre au besoin réel des journalistes/politiciens/juristes de disposer d'outils d'investigation (extraction, vérification, corrélation) et de représentation de l'information (synthèse, 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.***

— Mise en perspective de l'information par la création de représentations visuelles et de tableaux de bord.

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
3. la valorisation des informations proposée par une IHM (cf. Figure 2)

Le système que nous proposons extrait, analyse et met automatiquement les informations dans des tableaux de bord (cf. Figure 3). Il construit des indicateurs pour les destinataires qui peuvent également visualiser l'évolution dynamique de l'information gérée par un système multi-agents. Actuellement, plutôt que d'utiliser PCA ou tSNE [VH08] pour visualiser nos documents dans un espace 2D réduit, nous avons opté pour un système multi-agents "d'attrac-

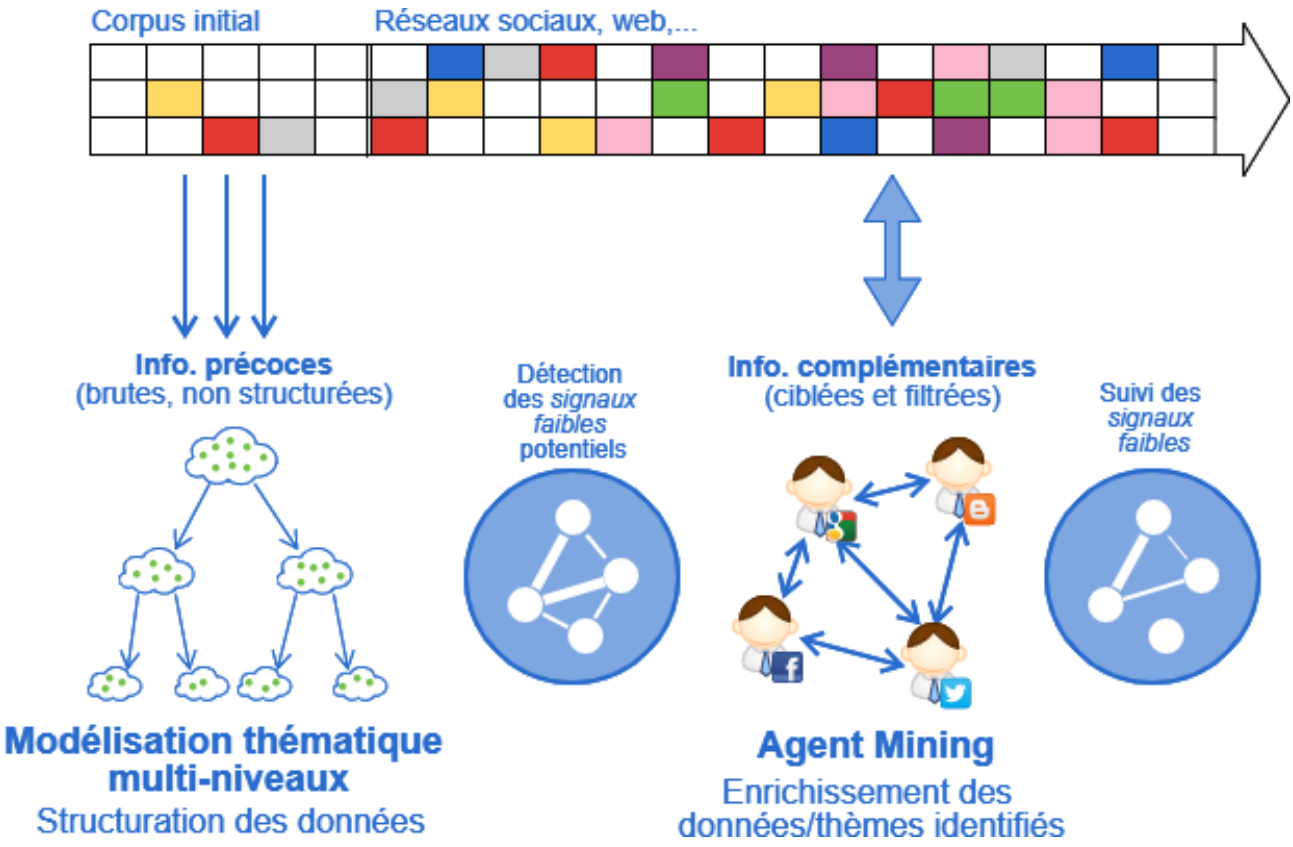


FIGURE 2 – ***Stratégie de détection des*** signaux faibles. ***Elle passe par l'analyse de l'information précoce apportée par un lanceur d'alerte et l'extraction des collections de mots associées aux Thèmes découverts. Ces informations permettent ensuite de mieux cibler la phase de*** data mining ***pour la détection des*** signaux faibles. ***Chaque rectangle de couleur représente une source d'information (ex : documents) exploitée.***

tion/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).

Sélection des *signaux faibles*

L'exemple de plateforme d'investigation pour le data journalisme que nous avons montré sur la figure 1 montre des acteurs interagissants entre eux par le biais d'outils et de processus sécurisés. Notre étude se positionne plus particulièrement sur la détection de signaux précurseurs dont la présence contiguë dans un espace de temps 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

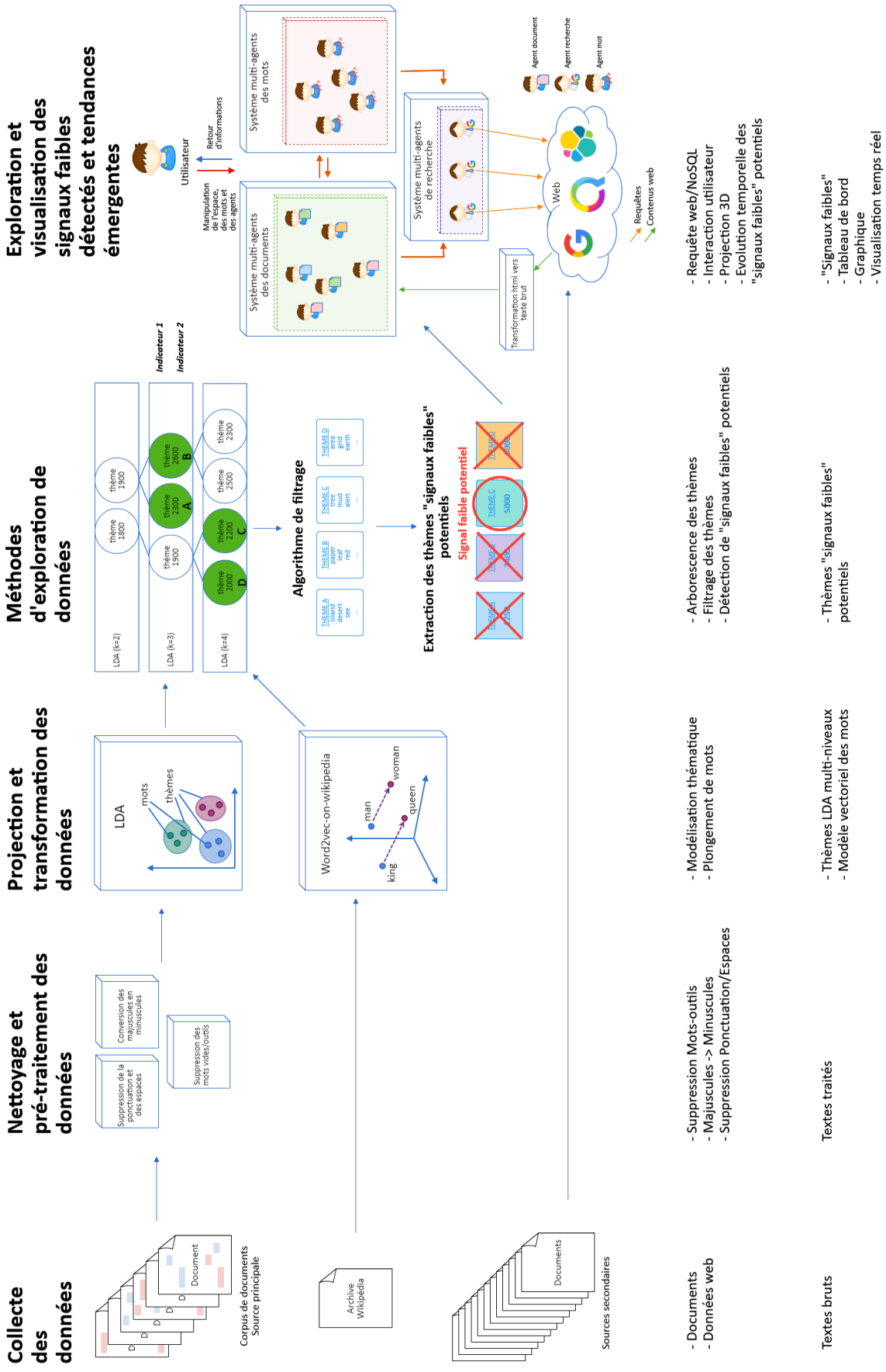


FIGURE 3 – *Détail des différentes étapes de l'extraction de connaissances/fouilles de données dans notre chaine de traitement ainsi que des propriétés et résultats obtenus.*

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.

Les travaux présentés répondent à une problèmatique de traitement automatisé de corpus de documents similaire à l'exploitation de données et aux investigations numériques réalisés dans le journalisme de données. Nous proposons également une nouvelle forme de visualisation interactive de données.

La diversité des sources de données présentée dans les publications scientifiques du domaine de notre étude montre la difficulté de proposer une approche uniforme [AZK14, TSV14]. L'investigation numérique dans un domaine spécifique nécessite la connaissance de termes pertinents à celui-ci [GU10, TSV14]. La recherche de ces termes peut s'effectuer à partir d'études extérieures ou être apportée *a priori* par des experts du domaine [GCPP05, MdFdA+14, HZM+15]. Le risque dans ce dernier cas est l'introduction de biais lors des phases suivantes de collecte des données [PVO13]. Nous proposons une solution s'appuyant sur des techniques de regroupement et de plongement de mots nécessitant une intervention moindre de l'expert du domaine.

L'exploration de données à partir d'un corpus traitant divers sujets est aussi un challenge. La croissance continue de production d'information et leur nature hétérogène complexifient l'extraction d'informations. La présence de documents, dont le contenu est non pertinent, empêche de distinguer les signaux précurseurs de faits ou d'événements et conduit à des difficultés d'interprétation. La solution que nous présentons, s'appuyant sur notre définition du *signal faible* (que nous donnerons plus loin), permet d'identifier/révéler l'information pertinente, représentée par une collection de mots présents dans les documents, lors de l'étape de modélisation thématique multi-niveaux (cf. Chapitre 2). Ensuite, lors de l'étape *agent mining* décrite dans le chapitre 3, nous utilisons l'enrichissement des données pour confirmer les informations trouvées et identifier un potentiel *signal faible*.

Comme nous le verrons au chapitre 1, un grand nombre de travaux de la littérature considère les corpus de données comme statique, et par conséquent ne proposent pas de procédures de mise à jour des collections de mots obtenus. Le corpus est figé, l'intégration de documents supplémentaires nécessite alors de procéder à une nouvelle analyse. Ces approchent ne permettent pas d'affiner les premiers résultats obtenus à partir de documents nouvellement détectés/apportés/produits. La fréquence élevée de production de données incite à proposer des systèmes de mise à jour constante des corpus [GU10, BXMH15, LCA15]. Dans nos travaux, nous intégrons une solution de veille à partir des informations trouvées dans les corpus initiaux et effectuons un suivi pour étudier leur évolution.

L'utilisation de techniques de visualisation des données est largement répandue en data journalisme. Ces graphiques ne permettent pas de forte interaction de l'utilisateur avec les données. Pour impliquer plus fortement l'utilisateur lors des phases d'analyse de données, la

mise en œuvre de tableaux de bord, d'indicateurs et de graphiques dynamiques permettent d'offrir des possibilités de retour de pertinence ainsi que de renforcement des choix.

Proposition d'une définition d'un *signal faible*

Dans le contexte plus particulier du monde de l'entreprise, des travaux tels que ceux de Ansoff [Ans75], ont démontré la nécessité de détecter le plus tôt possible les changements de contexte/condition/paradigme. La littérature propose deux formes courantes de changement : le "*signal faible*" et la "*tendance*". Celle qui nous intéresse tout particulièrement est le *signal faible* qui représente des signes précurseurs de changements critiques.

Nous proposons une définition du *signal faible* sur laquelle s'appuie notre système d'analyse de documents :

Definition. Un *signal faible* est caractérisé par un faible nombre de mots par document et présent 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).

Toutes les approches de modélisation thématique souffrent de la même difficulté : le nombre de thèmes obtenus ne correspond en général qu'à un optimum local. Même si les approches deviennent de plus en plus robustes, notamment grâce aux méthodes construites autour des processus de Dirichlet, la détermination du nombre de thèmes reste sensible à la structuration des observations et à l'information *a priori* disponible.

Nous proposons dans cette étude une approche méthodologique 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.

Nous prônons l'utilisation d'une approche conjointe : modélisation 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. 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.

Ce mémoire est organisé comme suit : tout d'abord, dans le chapitre 1, 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 et une *tendance*. Nous faisons une rétrospective selon plusieurs aspects des articles majeurs de la littérature. Nous concluons ce chapitre avec notre positionnement, présentons une définition suite à l'étude de la littérature, et proposons notre approche conjointe (globale et contextuelle) dont le but est de s'appuyer sur cette définition pour mettre en évidence le *signal faible*. Dans le chapitre 2, nous détaillons la première partie de la solution qui consiste en l'utilisation des méthodes de modélisation thématique (classification) et de plongement de mots qui sont toutes les deux impliquées. Cette partie présente l'une de nos contributions : une solution "*LDA* [3] augmentée par *Word2Vec*", appelée modélisation thématique multi-niveaux, pour la détection de *signaux faibles* potentiels. Nous présentons également les résultats obtenus, dans un premier temps, sur un corpus artificiel, puis dans un second temps sur un corpus de données réelles. Dans le chapitre 3, nous présentons la seconde partie de la solution, celle relative au domaine de l'*agent mining* mettant en œuvre une combinaison d'algorithmes issus du *data mining* et des systèmes multi-agents. Nous proposons une contribution permettant d'effectuer un suivi de *signaux faibles* potentiels par la recherche de nouveaux contenus en ligne ainsi qu'une proposition de visualisation interactive permettant de gérer en temps réel les documents porteurs de *signaux faibles*. Enfin dans le chapitre 4, nous expérimentons l'ensemble de la chaine de traitement sur des bases de documents tirés du corpus des projets H2020 du site d'information de CORDIS ainsi qu'une analyse bibliograhique d'une base de données d'articles. Nous étudions les résultats de la solution qui montrent l'intérêt de cette approche et proposons des perspectives d'amélioration.

Contributions

Dans ces travaux, nous proposons une chaine (1) d'extraction à partir de corpus de documents, (2) d'analyse semi-automatisée et (3) de recherche au moyen de requêtes Web pour *in fine*, proposer des tableaux de bord décrivant les *signaux faibles* potentiels. Cette chaine de traitement, non limitée dans la taille du corpus pouvant être traité, repose sur deux approches :

- **approche statique** où une analyse semi-automatique de documents permet d'extraire des *signaux faibles* potentiels dans des thèmes au moyen d'une modélisation thématique multi-niveaux ;
- **approche dynamique** où les groupes de mots-clés présents dans les thèmes obtenus précédemment sont utilisés pour formuler des requêtes Web enrichissant le corpus par de nouveaux documents. Dans un espace 3D, des agents animés par des forces d'attraction/répulsion, représentant documents et mots, se déplacent. L'utilisateur peut intéragir avec les agents formulant ainsi des requêtes en figeant et déplacant des agents, réorganisant alors l'affichage des autres agents et lançant de nouvelles recherches Web.

Après une période de temps défini par l'utilisateur, une nouvelle itération de l'approche statique est relancée afin d'obtenir des nouveaux thèmes. Ces nouveaux résultats peuvent servir à

3. Latent Dirichlet Allocation

effectuer un nouveau suivi de l'évolution des nouveaux thèmes obtenus.

Au cours des approches statiques et dynamiques, plusieurs étapes de la chaine de traitement ne peuvent être automatisées. Nous nous appuyons alors sur des experts pour évaluer les résultats obtenus. Le recours à des experts du domaine permet d'effectuer un travail d'investigation rapide et efficace afin de détecter rapidement les bons signaux porteurs d'informations utiles. Les outils d'aide à l'investigation offrent des capacités pour traiter, analyser et hiérarchiser cette information hétérogène.

La figure 3 représente les différentes étapes de la chaine de traitement selon le processus d'extraction de connaissances. La figure 4 décrit ces étapes sous une autre forme. Les phases de modélisation thématique (topic modeling), plongement de mots (word embedding) et d'élagage seront présentées dans le chapitre 2 correspondant à l'approche statique. Les étapes de projection et de découverte seront présentées quant à elles dans le chapitre 3 par l'approche dynamique.

L'ensemble de la chaine de traitement est mise en œuvre dans le logiciel WILD. Il fournit un ensemble de services et composants à destination de différents utilisateurs tels que des lanceurs d'alertes et des journalistes dans un objectif de prise de décision. Le logiciel, modulable, offre un ensemble de paramètres ajustables afin d'obtenir des résultats cohérents et pertinents.

Approche statique

La première partie de notre chaine de traitement propose une solution d'analyse semi-automatique de documents appelée modélisation thématique multi-niveaux afin d'extraire des *signaux faibles* potentiels. Elle est réalisée à partir de thèmes. Nous utilisons des approches qui ont fait leurs preuves comme *LDA* et *Word2Vec* mais qui possèdent certaines limites. *LDA* 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. *LDA* détermine seulement les thèmes prédominants d'une base documentaire. *Word2Vec* 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. 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. 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*. Puisque l'objectif est la détection d'un thème relatif au *signal faible*, celui-ci est par définition disjoint des autres (unitaire et non relié sémantiquement aux autres thèmes i.e. à des paradigmes existants), et relativement orthogonale aux autres informations contenues dans le corpus. Ce thème *signal faible* doit être suivi par un expert du domaine pour déterminer sa pertinence.

L'approche conjointe que nous proposons repose sur l'utilisation de *LDA* standard et de *Word2Vec*. Cette approche détermine les thèmes prédominants mais ceux-ci ne répondent pas à la définition du *signal faible*. Il nous faut pour le détecter, d'autres critères qui seront détaillés par la suite : *tf-idf*, distance de Bhattacharyya et critère de cohérence *Word2Vec*.

Dans un premier temps, pour plusieurs valeurs du nombre de thèmes, nous appliquons *LDA* standard sur le corpus de documents. Puis dans un deuxième temps, grâce à un critère de ressemblance, nous construisons une arborescence des thèmes obtenus. Cette arborescence est finalement simplifiée et élaguée grâce au critère de cohérence *Word2Vec*. Seuls les thèmes cohérents au sens de notre définition des *signaux faibles* sont alors retenus. *LDA* permet seulement de déterminer les thématiques fortes d'une base documentaire. Pour déterminer les mots-clés saillants, nous ré-évaluons la pertinence des mots qui composent les thèmes au moyen de la méthode de pondération *tf-idf*. Notre démarche repose donc sur un ensemble de critères pour obtenir les thèmes pertinents. Ceux-ci sont alors mis en perspective avec le domaine d'application par un expert pour déterminer leur pertinence.

Dans cette approche statique que nous appelons modélisation thématique multi-niveaux, nous utilisons donc les critères suivants :

- ***LDA*** permet d'obtenir les mots-clés les plus pertinents pour la sémantique du thème ;
- ***Word2Vec*** donne un critère de cohérence des mots appartenant aux thèmes ;
- **la distance de Bhattacharyya** permet d'obtenir les thèmes les plus disjoints possibles par le calcul de liens de ressemblance entre thèmes ;
- ***tf-idf*** calcule les mots-clés saillants d'un thème pouvant représenter plusieurs *signaux faibles* potentiels.

Un travail d'extraction de la sémantique de chaque thème est nécessaire. Celui-ci doit être réalisé par des experts du domaine. Néanmoins, la qualité des thèmes obtenus et des *signaux faibles* potentiels qu'ils contiennent est dépendant du nombre de documents. Un faible nombre de documents complexifie la détection des *signaux faibles* potentiels.

L'approche statique présentée dans cette première partie de la chaine de traitement permet seulement de déterminer les *signaux faibles* potentiels présents dans un corpus de documents. L'étude de leur évolution, au moyen de l'approche dynamique, permet de confirmer ou d'infirmer les *signaux faibles* potentiels détectés. Cette deuxième partie de notre chaine de traitement correspond à l'approche dynamique. Son objectif est de prendre en compte l'évolution temporelle de l'information.

Approche dynamique

La recherche de *signaux faibles* nécessite un suivi temporel. La première partie de ces travaux présente la détection de *signaux faibles* potentiels à partir d'une modélisation thématique multi-niveaux statique. Ces signaux nécessitent d'être corrélés à un contexte d'information plus large.

Pour cela, des requêtes sur des moteurs de recherche vont enrichir le corpus initial. Cette solution de veille d'information offre ainsi une capacité de suivi des *signaux faibles* potentiels. Celle-ci est supervisée par un expert qui détermine si l'évolution des thèmes met en évidence un *signal faible* précédemment détecté. Au moyen d'une interface sous Unity, des utilisateurs peuvent intéragir dans le système par des intéractions simples et intuitives, et visualiser l'évolution des thèmes.

Nous proposons ainsi une approche d'*agent mining* combinant *data mining* et système multi-agents. Les documents et mots-clés, obtenus durant l'approche statique, sont représentés dans un espace 3D où des agents, animés par des forces d'attraction/répulsion représentant ces documents et ces mots, se déplacent. Les thèmes et mots-clés saillants, obtenus durant l'approche statique de notre chaine de traitement, sont utilisés pour lancer des requêtes sur des moteurs de recherche. Les pages Web résultats de ces recherches sont ajoutées au système multi-agents sous la forme de nouveaux agents "document" afin d'enrichir le corpus. L'utilisateur peut intéragir avec les agents formulant ainsi des requêtes en figeant et déplacant des agents, réorganisant alors l'affichage des autres agents et lançant de nouvelles recherches Web à partir des mots-clés des documents.

Cet aspect dynamique permet de suivre l'évolution des mots-clés et les résultats des recherche de documents associés afin de déterminer si ceux-ci sont nombreux, quantitativement et qualitativement pertinents. Selon le nombre de documents reçus, il est alors possible de déterminer si un *signal faible* potentiel devient *signal fort*.

Expérimentations

Les tests réalisés dans cette thèse sur la première partie de la chaine de traitement dans un premier temps, puis sur l'ensemble de la chaine dans un second, ont permis une validation des concepts ainsi qu'une étude de pertinence sur des jeux de données réelles. Nous avons donc travaillé essentiellement sur deux bases de données :

— une base de données bibliographiques. Notre recherche porte sur l'évaluation de notre algorithme sur une base documentaire composée de l'état de l'art étudié par Mühlroth afin d'extraire éventuellement de nouveaux domaines non mis en avant par cet expert du domaine des *signaux faibles* et *tendances* émergentes.

— une base de documents scientifiques composée des résumés de projets H2020 à travers laquelle nous cherchons à déterminer les thèmes scientifiques actuellement étudiés au niveau européen.

Projets H2020

Nous utilisons la base de données des projets H2020 sur la période 2014-2021. Elle provient de l'Open Data du service d'information sur la recherche et le développement (CORDIS) de l'Union européenne (dump du 06/05/2020). Les sujets sont regroupés par année pour évaluer l'évolution des thèmes de recherche sur les 8 années. Nous utilisons les premières années comme

corpus initial, puis nous étudions l'évolution des thèmes et des mots-clés associés sur les années suivantes.

Les résultats ont donné des thèmes cohérents. Parmi ceux-ci, deux thèmes portant sur le domaine de la protection environnementale et des nouveaux matériaux se sont démarqués. Un troisième thème relatif au domaine médical et santé a permis de faire émerger deux thèmes lors de l'itération suivante portant sur le secteur médical, pour l'un, et la biologie cellulaire, pour l'autre.

Analyse bibliographique des articles étudiés par Mühlroth

Cette base documentaire est composée des résumés d'articles présents dans l'état de l'art étudié par Mühlroth. Nous utilisons les articles cités par l'auteur comme corpus initial de notre chaine de traitement. Nous avons ensuite récupéré les articles sur la période d'avril 2017 à septembre 2020 avec la même requête utilisée par Mühlroth. Les résumés de ces derniers sont utilisés pour le suivi temporel.

Nous avons évalué qu'avec un seuil autour de 14% de documents relatifs à un *signal faible*, nous obtenons des résultats intéressants de détection de celui-ci. Ce seuil cependant assez élevé est dû à la nature des documents qui sont de courts résumés dont la terminologie est plutôt générale. L'expérimentation sur des documents entiers donnerait certainement de meilleurs résultats.

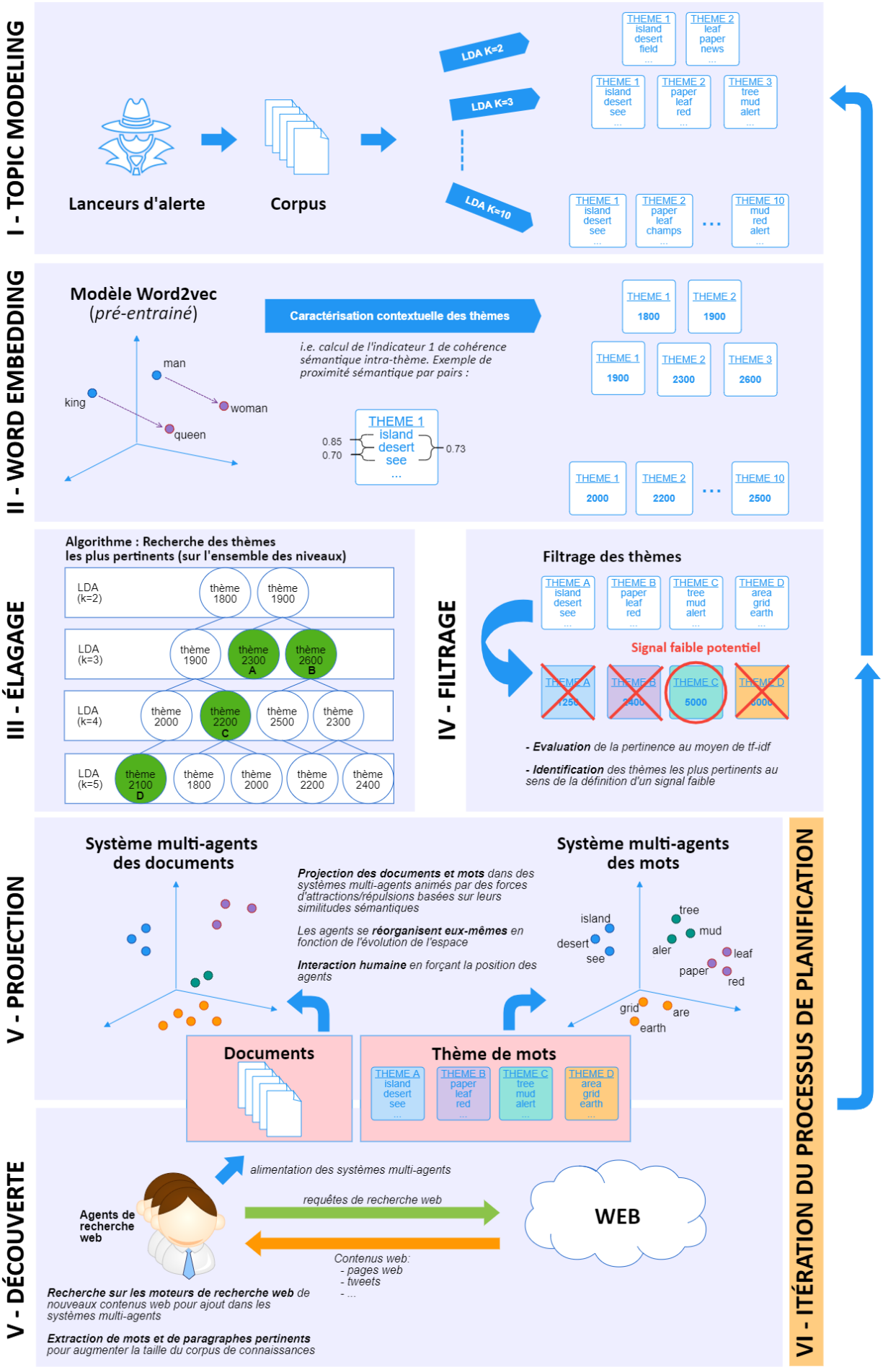


FIGURE 4 – ***Les différentes étapes pour extraire les*** signaux faibles

Chapitre 1

Définitions et Positionnement Scientifique

Sommaire

Contexte général et positionnement 5
Sélection des *signaux faibles* 8
Proposition d'une définition d'un *signal faible* 11
Contributions 12

1.1 Introduction

Dans ce contexte de croissance massive de la quantité d'information, la plannification stratégique et les actions managérialles sont devenues complexes. L'environnement des entreprises est soumis à des changements constants et à des prises de décisions nombreuses. De nouveaux facteurs, souvent inconnus, défient les entreprises, qui ne peuvent s'appuyer le plus souvent que sur de l'information incomplète et asynchrone [RB12]. Les quantités importantes de données disponibles nécessitent le développement d'outils toujours plus performants capables d'analyser les dynamiques en jeu [Hil08]. Ils doivent permettre d'exploiter les informations et extraire des connaissances à destination des décideurs sur les facteurs extérieurs à l'entreprise qui ont une influence directe ou indirecte sur elle [KvdG14].

Ansoff [Ans75] a souligné la nécessité de détecter le plus tôt possible, grâce à des outils d'analyse, les changements dans l'environnement des entreprises dans un objectif de planification stratégique. Ces changements peuvent se présenter sous la forme de discontinuités inattendues ou par l'émergence de *tendance* changeant fondamentalement la technologie et la société [Ans75, Kuo10]. Coffman [Cof97a] a proposé une définition plus précise de ce changement : une source qui affecte l'environnement des entreprises et ses activités et dont le caractère inattendu pour le récepteur potentiel est difficile à définir en raison des autres signaux et du bruit. Saritas et Smith [SS11] définissent quant à eux les deux formes courantes de ces changements :

— **"Signal faible"** : "Premiers signes de changement possible mais non confirmés qui peuvent devenir plus tard des indicateurs plus significatifs de forces critiques"
— **"Tendance"** : "Signes précurseurs qui découlent de changements et d'innovation largement généralisables"

Elles sont généralement considérées comme le début et la fin, respectivement, de l'évolution du *signal faible* précurseurs d'une *tendance* généralisée [Hil08, Kuo10, SS11]. Les *signaux faibles* sont les précurseurs des événements futurs et les *tendances* définissent les changements technologiques, politiques, économiques et sociétaux. Des exemples typiques de *signaux faibles* et *tendances* sont associés aux développements technologiques, aux changements démographiques, aux nouveaux acteurs, aux changements environnementaux, etc. [DS05]. Dans la section suivante, nous détaillons les différents aspects spécifiques de ce domaine de recherche et effectuons une analyse documentaire visant à présenter les idées et les connaissances de travaux antérieurs.

1.2 Angles d'analyse utilisés

Pour son étude portant sur 91 travaux de recherche allant des années 1997 à 2017, dans le domaine de la détection des *signaux faibles* et *tendances*, Mühlroth [MG18] a proposé 4 angles d'approches différents :

— **Selon les objectifs de détection**. La littérature montre la nécessité de détecter les deux formes courantes de changement que sont les *signaux faibles*, par la détection précoce de discontinuités, et les *tendances* émergentes signes de changement [Ans75, Kuo10]. On recherche ici à distinguer les articles selon ces deux objectifs : détection de *signaux faibles* et détection de *tendances* émergentes.
— **Selon le type de source de données**. Les types de source de données exploitées se voient détaillés selon une approche de codage inductif [Ber00] définissant des catégories.
— **Selon les domaines d'analyse**. Cet aspect est étudié selon la classification PEST (aussi appelée classification STEP) [Hil08, CN09]. Dans cette classification, les documents sont classés selon différents domaines (politiques, économiques, sociaux et technologiques). Combiner les domaines d'analyse aux objectifs de détection, il est alors possible de regrouper les études en 8 catégories.
— **Selon les techniques de fouille de données/extraction de connaissances**. Cet aspect est composé de 2 sous-aspects analysant les techniques de fouille de données utilisées :
  — **Approche de fouille de données/extraction de connaissances**. Cette classification permet de différencier les approches fouille de textes et fouille du web (*text mining*) pour l'analyse de modèles de données textuelles non struturées, l'analyse bibliométrique construite sur des modèles de données structurées telles que des métadonnées, citations ou des bases de données pré-structurées, et une approche combinatoire des deux.

— **Processus de fouille de données/extraction de connaissances**. Dans cet aspect, 5 étapes sont généralement considérées dans la chaine de traitement du processus : collecte des données, nettoyage et pré-traitement des données, projection et transformation des données, puis exploration de données (*data mining*) et enfin visualisation. Les approches méthodologiques employées sont regroupées selon la classification suggérée par Fayyad [FPSS96] :
  — **détection d'anomalies**. Identification d'enregistrements de données inhabituelles susceptibles d'être intéressantes ou d'erreurs de données qui nécessitent une analyse approfondie.
  — **regroupement (*Clustering*)**. Recherche de groupes ou de structures selon le degré de "similarités" des données, sans utiliser de structure *a priori*.
  — **classification**. Généralisation de la structuration des données appliquée à de nouvelles données (Par exemple, un programme de courrier électronique qui trie entre "légitime" et "spam" [Faw03] ou la détection de "fake news" sur les réseaux sociaux [SSW+17])
  — **groupement par similitude / règles d'associations**. Recherche de relation entre des variables (par exemple, un supermarché peut déterminer les habitudes d'achats des clients par l'apprentissage de règles d'association sur les fréquences d'achat des produits [AIS93])
  — **régression**. Estimation des relations entre une variable dépendante et une ou plusieurs variables indépendantes par l'apprentissage d'une fonction qui modélise les données avec le moins d'erreurs possibles.
  — **résumé automatique de texte**. Description compacte d'un sous-ensemble de données qui représente les informations importantes ou pertinentes au sein du contenu original.

1.2.1 Analyse selon les objectifs de détection

Dans l'état de l'art proposé par Mühlroth [MG18], 91 travaux sont donc étudiés selon les deux objectifs de détection : *signaux faibles* et *tendances*. Certains travaux fondateurs n'apparaissant cependant pas dans les résultats de Web of Science (WoS), site répertoriant toutes les revues dans le *Science Citation Index Expanded (SCIEXPANDED)* et sur lequel se sont appuyés les travaux de Mühlroth [MG18]. Pour ce dernier, l'objectif de ses travaux visait à découvrir l'état de l'art dans ce domaine de recherche plutôt qu'à retracer l'ensemble de son évolution. Plusieurs travaux fondateurs comme ceux d'Ansoff [Ans75] ne font alors pas partie de son étude. Ils sont seulement à la base d'outils d'analyses environnementales des entreprises pour les décideurs. Notre objectif est ici de proposer une étude historique des définitions d'un *signal faible* et d'une *tendance* montrant ainsi toute la complexité et par conséquent la difficulté de leurs détections (cf. Chronologies 1.1 et 1.2).

1.2.1.1 Signaux faibles

Une approche prospective dominante dans les entreprises et les organisations comme alternative ou complément à la planification stratégique est proposée par Ansoff dans les années 1970 et 1980 [Ans75, Ans80]. La planification stratégique des entreprises utilisée dans le cas d'un développement progressif n'est en effet pas suffisamment efficace dans des circonstances où le rythme du changement s'accélère. La planification stratégique nécessite des *signaux forts* telles que des informations suffisamment précoces et précises pour permettre une réponse adéquate. Pour Ansoff, les *signaux faibles* sont les premiers symptômes de discontinuités stratégiques : les symptômes d'un changement possible dans l'avenir, agissant comme des signes d'avertissement ou de nouvelles possibilités [Ans85]. Les réponses de l'entreprise doivent être adaptées en fonction de l'état des connaissances et vont, soit modifier la relation de l'entreprise avec l'environnement, soit modifier la dynamique et la structure interne de l'entreprise. Ansoff propose la définition des *signaux faibles* comme "symptômes d'un changement possible dans le futur" [Ans75, Ans80, Ans85].

Dans les années 1997, Coffman [Cof97a] a proposé une définition plus précise du *signal faible*. Selon Coffman, un *signal faible* est : (1) une idée ou une tendance qui affectera l'entreprise ou l'environnement commercial ; (2) une nouveauté surprenante du point de vue du récepteur du signal, bien que d'autres puissent également la percevoir ; (3) parfois difficile à repérer parmi d'autres bruits et signaux ; (4) une menace ou une opportunité pour une organisation ; (5) souvent raillé par les personnes qui "savent" ; (6) a généralement un temps de latence important avant de mûrir et de se généraliser ; et (7) représente donc une opportunité d'apprendre, de croître et d'évoluer [Cof97a, Cof97b, Cof97c, Cof97d, Cof97e].

Le concept de "wild card" est parfois utilisé comme synonyme de *signal faible*. Certains chercheurs soulignent la différence, la distinction entre le signe d'un phénomène et le phénomène lui-même : un *signal faible* est le signe d'un futur "wild card" [MCKoR04]. Pour Mendonça, les "wild cards" sont des phénomènes aux conséquences importantes et immédiates sur les parties prenantes de l'organisation lorsqu'ils se produisent [MCKoR04]. Cette distinction se trouvent aussi dans les textes originaux de Ansoff où la description des discontinuités stratégiques et des surprises stratégiques correspondent à l'idée de "wild card" sur lesquels les *signaux faibles* fournissent des informations précoces [Hil06].

La détection des *signaux faibles* doit parfois passer par une vision sur le monde extérieur. Day et Schoemaker [DS04, DS05] parlent de l'importance de scanner la périphérie pour détecter les *signaux faibles*. Ils expliquent qu'une vision périphérique tournée vers l'avenir permet aux personnes ouvertes d'esprit d'interpréter la dimension subjective des signes et de leur implication pour l'avenir [DS05].

D'autres travaux théoriques sur les concepts de *signaux faibles* ont été menés par Hiltunen [Hil06, Hil08]. Ces concepts de *signal faible* et "wild card" sont également distingués dans

les travaux de Hiltunen sur le déclenchement de la réflexion sur le futur et l'innovation dans les organisations [Hil07]. Dans sa terminologie, plus que simplement de futur "wild card", les *signaux faibles* peuvent indiquer de nombreux types de changements progressifs [Hil06]. Il définit le *signal faible* dans un espace tridimensionnel : (1) "signal" qui correspond à sa visibilité, (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 [Hil08]. Hiltunen [Hil08] suggère un nouveau terme pour *signal faible* : "signe futur".

Saritas [SS11] définit dans ses travaux la terminologie pour les *signaux faibles* comme "les premiers signes de changements possibles mais non confirmés" représentant les premiers signes de *tendances* futures.

On peut distinguer les *signaux forts* et les *signaux faibles*. Le *signal fort*, au contraire du *signal faible*, a un impact déjà présent sur une cible et agit également sur son avenir [MCC12]. Les *signaux faibles* peuvent être le signe d'un habitat inconnu que les acteurs peuvent involontairement ignorer, ou refuser d'accepter comme crédible, ou même ne pas s'y intéresser. Le développement de combinaisons équilibrées de capacités analytiques et sociales permet une meilleure compréhension des *signaux faibles* qui sont essentiels à toute entreprise stratégique innovante [MCC12].

Les travaux de Hiltunen [Hil08] propose la mise en œuvre d'un cadre subjectif et qualitatif nécessitant beaucoup d'argent et de temps pour que des experts puissent rechercher un *signal faible*. Cela rend difficile l'identification rapide de *signe futur*. La recherche de *signal faible* s'appuyant sur la subjectivité présente également une vulnérabilité en terme de garantie d'objectivité et augmente la difficulté d'améliorer la fiabilité du *signal faible*. Pour surmonter ces limites, Yoon [Yoo12] 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 de ces mêmes mots-clés qui montre le degré de diffusion (DF : document frequency).

Holopainen [HT12] présente une méta-analyse du concept de *signal faible*. Dans ses travaux, il retrace l'ensemble des études successives de l'idée de base présentée par Ansoff [Ans75]. Il décrit le *signal faible* comme un concept pouvant être relié à d'autres concepts d'avenir comme les *signaux forts* et les *tendances*. Les études récentes, présentant des approches pour identifier, collecter et interpréter des *signaux faibles*, font évoluer le concept et augmentent son applicabilité. D'après Holopainen [HT12], les travaux d'Ansoff [Ans75], pionnier du domaine, sont encore largement repris dans la littérature récente. Holopainen décrit les futurs directions possibles dans le développement du domaine des *signaux faibles*.

Thorleuchter [TV13, TSV14, TV15] dans ses travaux, identifie des *signaux faibles* sur la base du sujet décrit par une hypothèse. Les mots utilisés pour formuler l'hypothèse sont pris en compte par des experts humains pour des requêtes de recherche Web. Ces dernières permettent de récupérer des documents dans lesquels des sections pertinentes sont utilisées pour un traitement ultérieur. Ce traitement composé d'un modèle thématique avec différents niveaux

de partitionnement est utilisé dans une approche de maximisation du *signal faible*. Ce dernier est identifié à partir de 2 caractéristiques : l'occurence de termes caractéristiques (tel que des termes techniques spécifiques) et les cooccurences fréquentes (où un terme technique apparait plus fréquemment avec un autre terme que ce à quoi on pourrait s'attendre). Les motifs sémantiques textuels répartis en *signaux faibles* et *signaux forts* sont analysés manuellement quant à leurs impacts sur l'hypothèse et sont présentés au décideur pour la prise de décision stratégique. Ces travaux étudient l'évolution des *signaux faibles* à partir de collections de documents, récupérées à des points successifs dans le temps, permettant de suivre leurs développements dans une stratégique d'alerte précoce [TSV14].

Plus récemment, les travaux de Mühlroth [MG18], Rousseau [RCK21] et Eulaerts pour le Centre commun de la recherche (JRC) [EJG+19] proposent des études d'articles organisant les idées et les connaissances existantes du concept de *signal faible*. Van Veen [vVR21] propose une définition unifiée d'un *signal faible* à partir de l'étude de 152 articles répartis en 4 domaines de recherche dont 68 proposant une description spécifique de *signal faible*. Il le définit comme "Une perception de phénomènes stratégiques détectés dans l'environnement ou créés lors d'interprétation, éloignés du cadre perceptif de référence".

1.2.1.2 Tendances

Pour la détection de *tendances* émergentes, les premiers travaux sont proposés par Feldman dans une approche de découverte de modèle et d'analyse de *tendances* à partir de documents textuels [FD95]. Dans ses travaux, le texte annoté avec un ensemble de concepts est organisé de manière hiérarchique et traité comme une distribution de probabilité. La recherche de concepts pertinents est réalisée par une comparaison de distributions de concepts sur des périodes adjacentes. Cette comparaison, entre distributions de concepts utilisant anciennes et nouvelles données, permet de mettre en évidence l'émergence de *tendances*.

Les applications de détection de *tendances* sont principalement divisées en deux catégories [KGP+04] : semi-automatique et entièrement automatique. Des travaux proposés par Porter [PD95] en 1995, portent sur un système semi-automatique reposant sur les entrées de l'utilisateur comme première étape dans le détection d'une *tendance* émergente. Le système vient fournir les preuves indiquant si un sujet saisi est émergent sous forme de rapports et d'indicateurs qui résument les preuves disponibles sur le sujet. D'autres systèmes semi-automatiques existent [BPK+01, RGP02, Che06].

A la différence des systèmes semi-automatisés, les systèmes automatisés prennent en charge un corpus et développent une liste de sujets émergents sans intervention humaine [DHJ+98, SA00, SJ00, PKM01]. L'évaluateur humain passe les sujets en revue ainsi que les preuves trouvées par le système pour déterminer les *tendances* véritablement émergentes. Durant le processus, l'utilisateur suit de manière intuitive l'avancée du traitement aux moyens d'éléments

1975-1985	Ansoff - Approche de planification stratégique, 1re définition du *signal faible* : "les premiers symptômes de discontinuités stratégiques : les symptômes d'un changement possible dans l'avenir, agissant comme des signes d'avertissement ou de nouvelles possibilités" [Ans75, Ans80, Ans85]
1997	Coffman - 2e définition du *signal faible* : "(1) une idée ou une tendance qui affectera l'entreprise ou l'environnement commercial ; (2) une nouveauté surprenante du point de vue du récepteur du signal, bien que d'autres puissent également la percevoir ; (3) parfois difficile à repérer parmi d'autres bruits et signaux ; (4) une menace ou une opportunité pour une organisation ; (5) souvent raillée par les personnes qui "savent" ; (6) a généralement un temps de latence important avant de mûrir et de se généraliser ; et (7) représente donc une opportunité d'apprendre, de croître et d'évoluer " [Cof97a, Cof97b, Cof97c, Cof97d, Cof97e]
2004 - 2005	Day - Importance de la vision périphérique dans la recherche de *signaux faibles* [DS04, DS05] Mendonça - Concept de "Wild card" : "phénomène aux conséquences importantes et immédiates sur les parties prenantes de l'organisation lorsqu'il se produit" [MCKoR04]
2006 - 2008	Hiltunen - Définition d'un espace tridimensionnel (signal, problème, interprétation) pour représenter un *signal faible* [Hil06, Hil07, Hil08]
2011	Saritas - 3e définition du *signal faible* : "les premiers signes de changements possibles mais non confirmés" [SS11]
2012	Holopainen - Méta-analyse de l'évolution des recherches sur le domaine du *signal faible* [HT12] Mendonca - 4e définition du *signal faible* : "le signe d'un habitat inconnu que les acteurs peuvent involontairement ignorer, ou refuser d'accepter comme crédible, ou même ne pas s'y intéresser" [MCC12] Yoon - Approche avec carte d'émergence et carte d'émission pour caractériser un *signal faible* [Yoo12]
2013 - 2015	Thorleuchter - Approche d'occurence de termes caractéristiques et de cooccurences fréquentes. Prise en compte de l'évolution temporelle du *signal faible* [TV13, TSV14, TV15]
2018	Mühlroth - Analyse de la littérature dans le domaine de la détection des *signaux faibles* et *tendances* provenant de WoS [MG18]
2019	Eulaerts (pour le JRC) - Approche par exploration de texte et indicateurs scientométriques pour la recherche de *signaux faibles* sur la littérature scientifique en science et technologies [EJG+19]
2021	Rousseau - Analyse de la littérature sur les méthodes et applications pour la détection et l'identification de *signaux faibles* dans des grands ensembles de données [RCK21] Van Veen - 5e définition du *signal faible* : "Une perception de phénomènes stratégiques détectés dans l'environnement ou créés lors d'interprétation, éloignés du cadre perceptif de référence" [vVR21]

TABLE *1.1* – ***Chronologie des événements marquants dans la définition des*** signaux faibles

visuels.

D'autres travaux menés par Lent [LAS97] en 1997 proposent un système d'identification de *tendance* dans les brevets américains par l'emploi de requêtes de modèles textuels séquentiels généralisés. Ce système permet d'identifier les *tendances* des documents textuels collectés sur une certaine période de temps.

Plusieurs années plus tard, les travaux de Tho [THF03] proposent un système technologique de Web Mining sur des publications de recherches scientifiques afin de détecter des *tendances* technologies dans les données. Elles sont alors définies comme des fonctions du temps évoluant dans une certaine direction qui nécessite une base de données suffisamment importante pour être détectées [Hil08, Kuo10]. Pour Saritas [SS11], les *tendances* sont définies comme des "facteur de changement qui découlent d'innovations largement généralisables" durant plusieurs années et dont la portée est mondiale. L'extrapolation de *tendances* vers l'avenir est une tâche hasardeuse en raison du risque d'ignorer les développements non linéaires, chevauchants et surprenants tels que les "wild cards", les chocs et autres discontinuités [Kuo10, SS11]. Mathioudakis [MK10] propose une méthode de détection et d'analyse en temps réel sur le réseau social Twitter dans lequel les utilisateurs peuvent intéragir.

Plus récemment, les travaux de Mühlroth [MG18] proposent une étude d'articles organisant les idées et les connaissances existantes du concept de *tendance*. Parmi les articles cités marquants, on trouve plusieurs études sur Twitter [LCB12, APM$^+$13, XZJ$^+$16] ainsi que sur les brevets [CL11, TWTDT11].

1.2.1.3 Evolution du domaine et terminologie

Entre les années 2005 et 2016, ce domaine de recherche a connu un regain d'activité en terme de nombre de publications. Ces résultats montrent que le domaine de recherche de la détection des *signaux faibles* et des *tendances* suscite une attention croissante. Celle-ci porte cependant principalement sur la détection des *tendances* émergentes alors que l'augmentation des publications sur la détection de *signaux faibles* reste quant à elle constante.

La distinction entre *signaux faibles* et *tendances* émergentes est parfois ténue ou floue [EMLS14]. Dans certains travaux, les auteurs peuvent même ne pas faire de distinction entre *signaux faibles* et *tendances* ou utiliser une terminologie non précise (par exemple l'utilisation du terme "topic" (sujet) sans effectuer pour autant de modélisation thématique "topic modeling". D'autres auteurs utilisent de manière confondue "thèmes" et "clusters" [CWL10, TS12, BXMH15].

1995	Feldman - Découverte de connaissances dans les bases de données textuelles [FD95] Porter - *TOA* : Analyse des opportunités technologiques [PD95]
1997	Lent - *PatentMiner* : Découverte de *tendances* dans des bases de données textuelles [LAS97]
1998	Détection de nouveaux événements [AP98, YPC98]
2000 - 2001	Swan - *TimeMines* : Construction de calendriers à l'aide de modèles statistiques d'utilisation des mots [SA00, SJ00]
2001	Bank - *CIMEL* : Apprentissage multimédia en ligne basé sur l'enquête collaborative constructive [BPK+01] Pottenger - *HDDI*^{TM} : Indexation dynamique hiérarchique distribuée [PKM01]
2002	*ThemeRiver*^{TM} : Visualisation des changements thématiques dans les grandes collections de documents [HHWN02]
2004	Kontostathis - Etat de l'art sur la détection des *tendances* émergentes dans l'extraction de données textuelles [KGP+04] Tho - Identification de *tendances* technologiques par approche Web Mining [THF03]
2006	Chen - *CiteSpace II* : Détection et visualisation des *tendances* émergentes et leurs changements dans la littérature scientifique [Che06]
2010	Kuosa - *FSSF* : Concept d'outil d'analyse de signes futurs [Kuo10] Mathioudakis - *TwitterMonitor* : Détection de *tendances* émergentes en temps réel sur Twitter [MK10]
2011	Curran - Etude de la convergence entre des disciplines scientifiques, des technologies ou des marchés. Application aux domaines des aliments fonctionnels et nutraceutiques / cosméceutique ainsi qu'aux technologies de l'information et de la communication (TIC) [CL11] Trappey - Etude du développement de la technologie RFID à partir de la base de données de brevets de l'Office d'Etat de la propriété intellectuelle de la République populaire de Chine [TWTDT11] Saritas - Définition de différentes formes de *tendances* et propositions d'évolutions futures [SS11]
2012	Lau - Détection et analyse de *tendances* sur Twitter par modélisation thématique [LCB12]
2013	Aiello - Détection de *tendances* sur Twitter a l'aide d'une méthode exploitant la distribution temporelle des concepts [APM+13]
2016	Xie - *TopicSketch* : Détection en temps réel de sujets à l'aide de modèles thématiques basés sur des croquis [XZJ+16]
2018	Mühlroth - Analyse de la littérature dans le domaine de la détection des *signaux faibles* et *tendances* provenant de WoS [MG18]

TABLE *1.2* – ***Chronologie des événements marquants pour les*** tendances

1.2.2 Analyse selon la source de données

Mühlroth [MG18] a référencé 5 catégories de sources de données exploitées dans les articles de la littérature. Les sources utilisées sont :

— les publications scientifiques, dans 29 articles ;
— les brevets, dans 34 articles ;
— les sources webs, dans 14 articles ;
— les réseaux sociaux, dans 23 articles ;
— autres sources, dans 2 articles ;

Certains travaux utilisent conjoitement plusieurs sources de données.

Les publications scientifiques collectées comme sources de données proviennent de WoS [dCdSF06, GT12, WQZ+15], Scopus [WHM09], the National Digital Science Library of Korea [KHJJ12] et de plusieurs journaux spécialisés [MZ05, WM06, BEG09]. Ces travaux recherchent dans les articles de journaux ou de conférences les *tendances* émergentes comme par exemple celle des nanotechnologies [dCdSF06] ou mettent en application des méthodes d'aide à la prévision et à l'identification d'activités d'innovation [WQZ+15]. Les sources scientifiques portant sur un large panel de domaine, les outils d'identification et d'études d'évolution sont des facteurs clés de prise de décisions dans tous les domaines, aussi bien technologiques que scientifiques [WM06, BEG09].

Un grand nombre de sources spécifiques aux brevets existent parmi lesquelles l'office européen des brevets (European Patent Office) [CWL10, VBV10, Cav16], l'organisation mondiale de la propriété intellectuelle [VBV10], l'office des brevets du Japon [VBV10], l'index mondial des brevets Derwent (Derwent World Patents Index) [WSLS10, HZM+15, MP15] ou encore le bureau américain des brevets et des marques de commerce (United States Patent and Trademark Office) [GM12, JPJ12a, GJS13]. Ces études recherchent au moyen de l'analyse bibliométrique de brevets et l'analyse des réseaux de brevets, les *tendances* technologiques (e.g. utilisation d'émetteurs en nanotube de carbone pour le domaine de la technologie FED (field emission display) [CWL10]). Ces brevets sont l'objet d'études lors de fusions technologiques dans des collaborations interentreprises, aidant les décideurs politiques et les gestionnaires à prévoir l'émergence des nouveaux domaines technologiques [Cav16]. Ces études permettent également une surveillance technologique en réduisant les coups et l'incertitude sur des inventions innovantes (e.g. l'industrie automobile [GM12]).

Parmi les sources de données Web, on trouve les sites d'information ainsi que les groupes de discussion [NRC06, LSLL09, KKB+13], les sites et les blogs [GU10, VBV10, TSV14] et les flux RSS [DFVL14]. Dans le cadre d'études de l'opinion des gens sur un produit ou une entreprise, les traitements d'articles d'actualité, de billets de blogs, sites d'évaluations et tweets permettent de déterminer les opportunités et risques pour des professionnels de la communication et leur organisation [GU10]. On trouve également l'étude des *signaux faibles* au moyen de séquences

de partitionnement (*clustering*) mesurées à des moments successifs dans le temps pour la prise de décision d'organisation [TSV14].

Les réseaux sociaux font partie des nouvelles sources de données aussi diverses que variées qui incluent aussi bien les marque-pages sociaux (*social bookmarking*) pour stocker, classer, chercher et partager des liens entre internautes [WZB10], les forums et communautés thématiques en ligne [LZL+13], Twitter [AOGS13, RRSZ14, SKJ14], Tencent QQ (une plateforme de messagerie instantanée chinoise) [WLT+15] Wikipedia [KTKK15], que les métadonnées d'images de Flickr [BXMH15]. Des sujets brûlants, comme le secteur de la santé, sur les thématiques du cancer du poumon, du cancer du sein et du diabète font l'objet d'étude à partir de communautés de santé en ligne dans lesquelles les internautes partagent leurs expériences et des connaissances sur les soins de santé [LZL+13]. Différentes plateformes en ligne comme Twitter, New York Times et Flickr permettent l'élaboration de recherche multimodale d'enrichissement d'information pour une solution de détection et d'élaboration de sujets émergents [BXMH15].

Des sources diverses peuvent être utilisées comme Moreira [MHCS15] qui utilise une base données de *signaux faibles* pré-remplie manuellement par des experts du domaine sous forme de phrases synthétiques qui se révèlent être pertinentes pour une analyse d'évolution du domaine. Ena [EMSS16] utilise plusieurs bases de données contenant des projets prospectifs, ceux notamment choisis par la Commission Européenne, ainsi qu'un agrégateur de journaux et de magazines.

1.2.3 Analyse selon les domaines d'analyse

Dans cette section, nous souhaitons distinguer les approches selon 4 domaines d'analyse et en prenant en compte les deux objectifs de détection : *signaux faibles* et *tendances*.

1.2.3.1 *Signaux faibles* dans le secteur économique

La recherche des *signaux faibles* dans le secteur économique se concentre sur l'étude des changements macro et micro économiques dans l'environnement des entreprises. Ces changements interviennent sur des domaines d'intérêt stratégiquement pertinents pour les entreprises, qui influent sur la stabilité économique [BI06] et mettent en évidence des annonces et événements économiques inattendus dans leur environnement [LSLL09].

1.2.3.2 *Signaux faibles* technologiques

Dans un objectif de plannification stratégique, la recherche et l'identification des signaux précoces et des changements dans l'environnement R&D sont les moteurs de l'extraction de

données dans le contexte technologique. Pour une entreprise, l'objectif principal n'est pas tant l'identification et le suivi des développements technologiques que l'extraction de l'information pertinente dans de grandes collections de documents pour la prise de décision dans les choix d'investissement R&D [VBV10]. L'identification des menaces ou des opportunités technologiques permet de soutenir la génération d'idées dans la compétition technologique toujours plus intense [YP07]. Les résultats sont aussi utilisés dans un but de plannification stratégique de mise en place d'une technologie [YK12]. Ils permettent par exemple, au regard d'un audit sur les compétences internes de l'entreprise, d'optimiser les investissements pour aquérir cette technologie émergente [VBV10]. L'évaluation de la pertinence aide les chercheurs, les praticiens et les entreprises dans leurs décisions de poursuivre ou négliger de nouveaux développements [GJS13].

1.2.3.3 *Signaux faibles* dans les secteurs politiques et sociaux

Ces deux domaines n'ont pas été présentés dans les articles analysés sur la période 1997 à 2017. Comme l'explique Mühlroth [MG18], la recherche de facteurs sociaux émergents sur des bases de données est perçue comme particulièrement complexe, car s'appuyant sur un flux constant de textes dont l'hétérogénéité est ingérable. La détection de *signaux faibles* par un système sur de telles données a de forte chance de produire un nombre de résultats proches du spam pour les utilisateurs [GU10, BXMH15, LCA15].

1.2.3.4 *Tendances* dans le secteur politique

L'étude des *tendances* politiques au travers des canaux de communication traditionnels(tels que les journaux et la télévision) ainsi que les canaux de communication en ligne (sites web, blogs et médias sociaux) au moyen de l'extraction de données peuvent aider à la détection des sujets et des opinions politiques [RRSZ14, SKJ14] ainsi que des injonctions et évènements politiques [MZ05, GV17] susceptibles d'influencer l'environnemnt des entreprises. La détection de *tendances* précoces, notamment aux moyens des réseaux sociaux, peut être utilisée pour de l'analyse des sentiments en étendant des bases de connaissances existantes (ontologies web ou réseaux sémantiques) [RRSZ14]. L'observation *a postériori* de données de réseaux sociaux permet d'étudier des comportements ainsi que l'influence et l'évolution de sujets controversés, plus particulièrement sur des groupes d'utilisateurs aux dispositions politiques similaires. Ainsi leurs connexions mutuelles occultent leurs visions du comportement de l'opinion dans le monde [SKJ14].

1.2.3.5 *Tendances* dans le secteur économique

La détection des *tendances* économiques liées aux facteurs macros et microéconomiques permettent d'analyser les changements aux moyens de données publiques disponibles. Ces changements peuvent par exemple porter sur des informations financières ou législatives à destination des entreprises [DCWX10] ainsi que sur l'analyse de modèles de convergence d'industries [PNML12, WRP+13, KLK+15] et d'observation des frontières, précédemment disjointes, entre les marchés et les industries. Des résultats récents démontrent cette convergence par une similarité sémantique croissante au fil du temps entre des brevets de domaines technologiques disjoints sur des secteurs tels que l'alimentation, les produits pharmaceutiques et les soins personnels [PNML12]. Cette convergence crée de nouvelles industries comme l'industrie de la nutrition médicale, rapprochement entre les industries de l'alimentation et de la pharmacie, dépassant même le stade de la convergence en matière de diffusion de l'information et actuellement en phase de consolidation de convergence industrielle [WRP+13]. La consolidation est déterminée par une co-classification des brevets et montre l'importance pour les entreprises situées à la frontière de ses domaines de suivre les développements technologiques.

1.2.3.6 *Tendances* dans le secteur social

Les évolutions du monde réel peuvent être détectées aux moyens de données provenant de sources sociales et agissant comme capteurs [APM+13]. Ces énormes quantités d'informations riches et opportunes sur des événements réels de toutes sortes proviennent majoritairement de médias sociaux et d'information en ligne et nécessitent de nouvelles techniques de pointe pour traiter des flux de données toujours plus complexes, denses, hétérogènes, condensées et synthétiques ou incomplètes. Dans un environnement social toujours plus en mouvement, réagir aux nouvelles *tendances* émergentes par leur détection est un facteur clé pour les entreprises [CCS13]. Cette détection en temps réel, à partir par exemple de fils de discussion émergents, nécessite de nouveaux outils comme l'analyse topologique des données, pour comprendre comment l'information se répand en exploitant la structure topologique et géométrique de données sous-jacentes. De multiples applications sont possibles :

— une entreprise peut effectuer une étude d'opinion sur l'entreprise et ses produits auprès de ses clients [GU10, LPZ15]. L'étude des traces numériques des intéractions entreprise/-consommateurs peut aider les chefs d'entreprise à mieux comprendre l'intérêt potentiel des consommateurs [LPZ15] ;
— des sujets liés à la santé apparaissent très souvent [LZL+13, PWY+13]. Par exemple, le suivi des termes les plus fréquemment utilisés dans les réseaux sociaux permet de détecter des changements de santé publique en termes de conditions de préoccupations de santé publique [PWY+13] ;
— l'étude des recommandations et des incidences des produits "à la mode" [FZYL14]. L'étude des relations multiples entre messages issues de réseaux sociaux, telles que

liens sémantiques et temporels, tags et hashtags, au moyen de méthodes d'extraction de mots-clés s'appuyant par exemple sur un arbre des suffixes, permet la détection de sujets reflétant des aspects de la vie quotidienne des personnes ;

— le suivi de discussions intensives [WLT+15]. Pour ne citer que lui, le microblogging, grâce aux métadonnées, témoignent, à travers des messages limités, des sujets "brulants" et jugés pertinents ;

1.2.3.7 *Tendances* dans le secteur technologique

L'étude de l'évolution dans le temps des *tendances* technologiques, à partir de documents sous forme numérique, peut permettre d'identifier le paysage de la R&D dans un domaine [EMSS16]. Le suivi des *tendances* technologiques, comme le nombre croissant de sources de données ou l'adoption d'approches de regroupement personnalisées pour identifier des *tendances*, nécessite des méthodologies systématiques de suivi comme par exemple l'analyse de données bibliométriques [EMSS16]. Les décisions d'investissement en R&D des décideurs en entreprise doivent prendre en compte les nouveaux développements technologiques en explorant de nouvelles opportunités commerciales, tout en se défendant des menaces éventuelles [SLH10, LJP11]. Les brevets sont une source importante de veille concurrentielle offrant un avantage stratégique pour les entreprises. De nouvelles approches sont développées pour extraire des changements de *tendances* sans intervention humaine comme par exemple l'étude sur deux périodes différentes des changements de *tendances* des brevets pour ensuite être évalués puis classés selon leur degré de changement [SLH10]. Les entreprises utilisent ses nouvelles méthodologies pour développer des produits et des stratégies technologiques innovants [NSY16]. Les *tendances* technologiques et les décisions d'investissement anciennement basées sur des connaissances et expériences intrinsèques [WCK10] font appel de nos jours à des techniques complexes d'exploration de données [WHM09, SLH10, CZZL15, NSY16]. La détection et le suivi temporel d'une technologie émergente ou d'un domaine technologique combiné à l'état actuel des connaissances et de la technologie indiquent l'évolution des développements futurs dans le but de prédire les trajectoires futures [HZM+15, JY15, MP15, WQZ+15]. La convergence des domaines technologiques, effaçant les frontières entre des disciplines scientifiques, peut être identifiée par l'extraction des données dans des modèles d'étude de convergence [CL11]. Cela fourni des indicateurs précieux pour les décideurs dans leurs actions de fusions et acquisitions dans des domaines tels que les technologies de l'information, de l'électronique grand public et des télécommunications. Des domaines comme le domaine des technologies de l'information et des communications (TIC) et le domaine de l'électronique ont fusionné en plusieurs secteurs tels que l'électronique imprimable, les téléphones intelligents et les étiquettes d'identification par radiofréquence (RFID) fournissant de nouvelles opportunités technologiques [Cav16].

1.2.4 Techniques de fouille de données/extraction de connaissances

Cette section est structurée selon les classes et sous-classes de techniques d'exploration de données décrites en introduction (section 1.2). Nous étudions plus particulièrement les techniques d'exploration des données employées. Nous proposons une description de l'état de l'art et une méta-analyse bibliographique effectuée grâce à notre système. Chaque classe et sous-classe des techniques d'exploration sont présentées sous ce prisme : d'abord une synthèse décrivant les méthodes utilisées dans la littérature, puis celle adoptée dans notre système pour cette méta-analyse bibliographique. L'objectif est triple : 1) comprendre, en étendant l'état de l'art de Mühlroth [1], les *tendances* actuelles des approches utilisées ; 2) montrer la pertinence de notre système, s'appuyant sur une base de données bibliographiques déjà référencée, pour rechercher de nouveaux documents pertinents, et 3) illustrer sur un premier exemple la méthodologie adoptée qui sera précisée dans les chapitres 2 et 3. D'autres résultats seront présentés dans le chapitre 4.

1.2.4.1 Processus de fouille de données/extraction de connaissances - Etat de l'art et méta-analyse bibliographique

Dans le but d'étudier plus en profondeur les 91 articles relevés par Mühlroth [MG18] sur la période 1997-2017 et l'étendre sur la période actuelle, nous utilisons notre chaine de traitement sur les résumés d'articles. Dans la suite de la présentation, et pour chacune des étapes, nous détaillons les traitements mis en œuvre (cf. Figure 1.1) et présentons les résultats de nos tests. Cette méta-analyse permet d'étudier le présence de *signaux faibles* potentiels au sein des résumés des articles pour pouvoir ensuite étudier leur évolution et les liens avec d'autres articles du domaine sortis après les travaux de Mühlroth [MG18].

1.2.4.1.1 Collecte des données

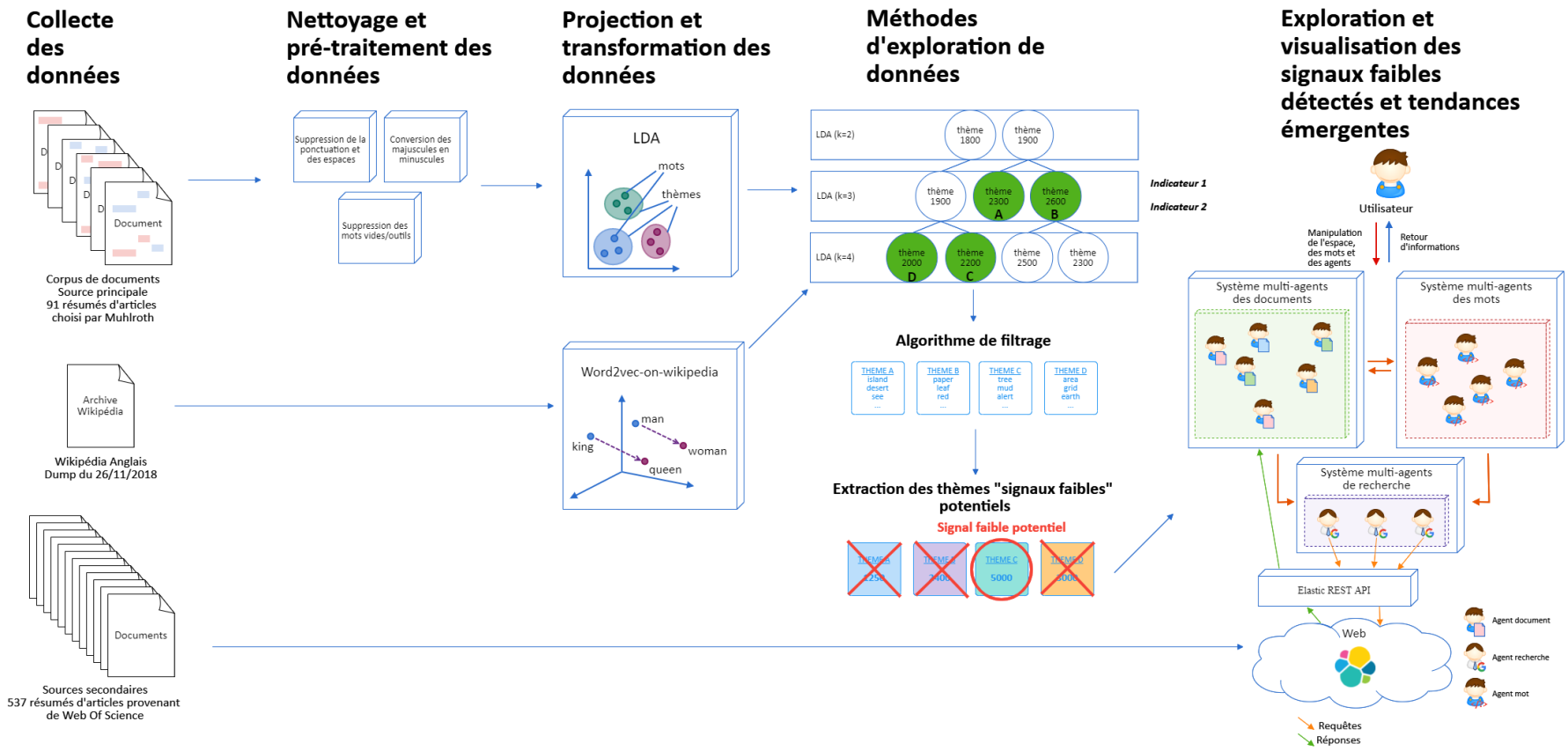
Approche générale - état de l'art

L'exploration de données doit être réalisée sur un échantillon de données extrait via des requêtes afin de se concentrer sur des contenus spécifiques pertinents [AZK14, TSV14]. La formulation

1. Les travaux de Mühlroth [MG18] reposant sur des articles issus de la période 1997-2017 décrivent des études utilisant des approches de type text mining, analyse bibliométrique ou conjointement les deux, la première approche étant majoritaire.

Le text mining est principalement utilisé pour des objectifs d'analyse se concentrant sur l'identification des facteurs politiques, économiques ou sociaux. La détection des *signaux faibles* et des *tendances* en R&D fait appel au text mining ou à l'analyse bibliométrique. Des exceptions existent comme l'article de Liu [LSLL09] qui détecte des *signaux faibles* économiques dans des pages web d'information au moyen d'une analyse conjointe.

L'analyse bibliométrique ne semble pas avoir été utilisée pour la détection des *signaux faibles* économiques et des *tendances* politiques. L'application conjointe de l'exploration de textes et de l'analyse bibliométrique a été particulièrement utilisée dans les domaines technologiques pour détecter les *tendances* et *signaux faibles* de nature technologique.



Propriétés	- Documents	- Suppression Mots-outils - Majuscules -> Minuscules - Suppression Ponctuation/Espaces	- Modélisation thématique - Plongement de mots	- Arborescence des thèmes - Filtrage des thèmes - Détection de "signaux faibles" potentiels	- Requête NoSQL (REST API) - Interaction utilisateur - Projection 3D - Evolution temporelle des "signaux faibles" potentiels
Résultats	Textes bruts	Textes traités	- Thèmes LDA multi-niveaux - Modèle vectoriel des mots	- Thèmes "signaux faibles" potentiels	- "Signaux faibles" - Tableau de bord - Graphique - Visualisation temps réel

FIGURE *1.1* – ***Détail des différentes étapes de l'extraction de connaissances/fouilles de données de notre chaine de traitement sur les articles de WoS ayant trait aux*** signaux faibles**.**

des requêtes peut être définie moyennant certaines hypothèses [TSV14]. La probabilité d'obtenir des résultats non pertinents augmente lors d'une collecte de données où les hypothèses ne sont pas vérifiées et l'ensemble des requêtes mal ou peu définies [GU10]. L'extraction de phrases et groupes de mots en fonction d'une occurence plus importante ainsi que l'étude de leur voisinage (mots ou groupe de mots adjacents) offrent des informations potentiellement pertinentes pour les décideurs d'entreprises sur les sujets émergents, par exemple lorsqu'un salarié d'Apple effectue une vieille technologique d'articles, de blogs et tweets sur les entreprises Apple, Microsoft, HP et Dell et leurs produits [GU10], ou l'utilisation d'un ensemble de mots comme filtre inital pour extraire les messages pertinents d'un réseau social comme Twitter [APM+13].
De manière évidente, une requête non suffisamment spécifique entraine un risque de résultats trop généraux. Cependant, l'inverse, une requête trop spécifique pose les risques d'un faible nombre de résultats ainsi que la perte de données potentiellement importante.
Les requêtes, construites à partir d'une proposition directe d'experts (ou sur la base d'hypothèses formulées) ou encore à partir d'un retour d'information sur la pertinence de premiers résultats, introduisent potentiellement un biais dans la collecte des données [PVO13]. Les résultats peuvent être incomplets et orientés de part la vision des experts sur leur domaine [GCPP05, MdFdA+14, HZM+15]. La recherche au moyen de termes liés aux domaines issus à la fois d'études antérieures et par les experts eux-mêmes sont des solutions régulièrement utilisées [MdFdA+14, HZM+15].

Les objectifs d'analyse diffèrent selon les domaines. Pour les domaines politiques et technologiques, l'objectif d'analyse repose généralement sur des requêtes formulées par les experts. Ces requêtes nécessairement spécifiques évitent l'obtention de résultats trop généraux. On peut citer la recherche de *tendances* politiques concernant des partis ou communautés [RRSZ14] ou des sujets de R&D hyper spécialisés [WCK10, TWTDT11] comme par exemple les nanotechnologies [dCdSF06]. Dans le premier cas, l'information à extraire doit être fortement ciblée. Dans le second cas, le langage utilisé est trop spécifique pour se passer de l'avis des experts.

La majorité des travaux ne met pas en œuvre de techniques de collecte de données permettant une mise à jour du corpus. Les données collectées sont alors considérées comme statiques et aucun autre document ne peut être inclus par la suite. L'étude porte donc sur l'état du corpus à un instant défini. L'intégration de documents supplémentaires par la suite nécessite alors de procéder à une nouvelle analyse sur l'ensemble du nouveau corpus et ne permet pas d'affiner les résultats précédents sur la base des documents récemment détectés ou d'évaluer leur évolution. Les travaux de Thorleuchter [TSV14] sont les seuls travaux trouvés par Mühlroth [MG18] qui exploitent des données web dans le domaine technologique et permettent de mettre à jour le corpus. Dans les approches partant sur les facteurs économiques, politiques et sociaux, les possibilités de mise à jour des corpus sont beaucoup plus répandues et se concentrent sur des sources web et données des médias sociaux. Les sources de données sont alors percues comme un flux constant de contenus textuels [GU10, BXMH15, LCA15] qui doivent être collectés à une fréquence élevée. Ceci explique les recherches connexes sur le développement de systèmes

de mise à jour régulier de corpus [GU10, LCA15]. Les travaux de Pinto [LCA15] utilisent le processus de Hawkes, qui est un modèle de diffusion de l'information. Ce dernier permet d'étudier la diffusion de sujets sur un réseau social afin de rechercher des "pics" dans l'intensité du processus et pouvoir déterminer les sujets susceptibles d'être *tendance* dans le futur. Goorha [GU10] propose la recherche de phrases décrites comme "explosives" à partir de noms de produits et d'entreprises dans un corpus de documents mis à jour continuellement. Une phrase est déterminée comme étant "explosive" si l'utilisation de cette dernière montre une augmentation spectaculaire durant différentes échelles de temps allant de 1 jour à 3 semaines.

Méta-analyse bibliographique - Mise en œuvre
Pour le corpus initial, nous utilisons les résumés des articles référencés par Mühlroth [MG18] sur la période de 1997 à 2017. Notre système s'appuie sur une modélisation thématique multi-niveaux qui est l'application conjointe d'une modélisation thématique et d'un plongement de mots. Pour ce dernier, nous utilisons l'approche *Word2Vec* et entrainons celui-ci sur l'ensemble des articles du Wikipédia Anglais (Dump du 26/11/2018).

Concernant la deuxième partie de notre chaine de traitement utilisant une méthode d'*agent mining*, nous adoptons la requête utilisée par Mühlroth [MG18] sur la base de données de WoS pour la période d'avril 2017 jusqu'à septembre 2020 (cf. Tableau 1.3). La requête permet d'obtenir 537 résultats. Les résumés de ses articles sont placés dans un index Elasticsearch [2] afin de simuler un moteur de recherche où des agents exécuteront les requêtes.

1.2.4.1.2 Nettoyage et pré-traitement des données

Approche générale - état de l'art
Les techniques de nettoyage et de pré-traitement permettent d'accroître la performance de l'analyse de données [APM+13, TV13]. Ces techniques ne sont souvent pas ou peu décrites, et probablement vues comme une étape qui ne nécessite pas d'explication par les auteurs car trivial ou sans intérêt.

Les techniques généralement utilisées sont le filtrage des caractères indésirables, de la ponctuation et des espaces [BOMOC14, LPZ15, WLT+15], la conversion des majuscules en minuscules [WM06, GWB11, TSV14], le tri [WCK10, LJP11, LZL+13], la suppression des mots vides/outils au sens sémantique [WCK10, YK12, LPZ15], la correction des erreurs typographiques [TSV14], le filtrage des caractéristiques à partir de la loi de Zipf ou ses variations [WM06, TSV14, HZM+15], le filtrage des caractéristiques s'appuyant sur trop ou trop peu de caractères [WM06, TSV14], la correspondance floue des mots [HZM+15] (combinaison de termes ayant des structures similaires sur la base d'un modèle commun tel que l'étymologie), le remplacement/l'agrégation des synonymes [APM+13, MHCS15] et la participation d'experts au nettoyage et au pré-traitement des données [GWB11, KKB+13, HZM+15].

2. `https://www.elastic.co`

Requête	TI=(("weak signal*" OR trend* OR technolog* OR topic* OR "research f*" OR "technolog* opportunit*" OR converg* OR fusion) AND (converg* OR detect* OR discov* OR emerg* OR evol* OR identif* OR mining OR monitor* OR scan* OR trac*)) AND TS=(((trend OR data OR text) AND mining) OR "trend detection" OR "technolog* forecast*" OR "technolog* intelligence" OR "technolog* opportunit*" OR "emerging topic*" OR "topic detection" OR "topic tracking")
Mots-clés	weak signal* trend* technolog* topic* research f* technolog* opportunit* converg* fusion detect* discov* emerg* evol* identif* mining monitor* scan* trac* data text trend detection technolog* forecast* technolog* intelligence emerging topic* topic detection topic tracking
Nombre de résultats	537
Période	Avril 2017 à Septembre 2020

TABLE *1.3* – ***Requête de recherche, mots-clés utilisés, nombre de résultats obtenus et période sur laquelle s'applique la recherche***

Il est possible également de ne pas souhaiter utiliser de techniques de filtrage et de prétraitement des données dans le but d'évaluer la robustesse d'algorithmes nouvellement introduits [MZ05]. Cataldi et al. [CCS13] filtrent les mots dans les statuts des utilisateurs de Twitter en adaptant les méthodes standards d'analyse de texte prenant en compte les techniques de type fréquence inverse. D'autres solutions ont été expérimentées comme la génération et la sélection de mots-clés pour des traitements ultérieurs au moyen de tâches itératives et manuelles [LKS+14] ou l'approche s'appuyant sur les anomalies dans un modèle pour détecter l'émergence de nouveaux sujets sur la base des relations réponse/mention dans les messages des réseaux sociaux [TTY14]. Certains auteurs n'évoquent pas de raison particulière à omettre l'étape de nettoyage et de prétraitement des données [SKT+11]. Il est justifié et nécessaire d'appliquer des mécanismes de filtrage et de nettoyage des données quand celles-ci sont de qualité inégale [WM12]. La tâche de prétraitement des données est parfois intégrée dans la tâche d'exploration de texte et à la projection spatiale vectorielle de mots clés, comme pour Jun et al. [JPJ12b] où ils n'effectuent aucune tâche de nettoyage et de prétraitement des données au sens propre du terme.

Méta-analyse bibliographique - Mise en œuvre

Dans le cadre de notre processus de traitement pour l'étape de modélisation thématique multi-niveaux, nous utilisons deux outils : *LDA* et *Word2Vec* (cf. Figure 1.2). *Word2Vec* utilise un modèle ne nécessitant pas de traitement supplémentaire des données. Pour *LDA*, il est impératif d'utiliser des techniques de nettoyage poussées telles que (cf. Figure 1.2) :

— la suppresion de la ponctuation et des espaces,
— la suppression des mots vides/outils au sens sémantique,
— la conversion des majuscules en minuscules

La suppression des mots vides/outils est réalisée au moyen de listes provenant de Github[3].

1.2.4.1.3 Projection et transformation des données

Approche générale - état de l'art

Dans la littérature, deux approches principales sont utilisées pour effectuer la projection et la transformation des données :

— **Les approches automatisées** projettent et transforment les données dans un format structuré sans aucune intervention humaine.
— **Les approches assistées** par des experts nécessitent une assistance humaine lors d'au moins une étape.

A la différence des autres domaines d'analyse qui reposent majoritairement sur des approches entièrement automatisées, la détection des *signaux faibles* technologiques utilise principalement des techniques de projection et de transformations des données reposant sur des experts. Parmi les travaux qui ont recours à une assitance humaine, Yoon et Park [YP07] créent de manière itérative, en s'appuyant sur des experts du domaine, une structure prédéfinie appelée matrice morphologique, utilisée comme base de leur analyse pour identifier la forme morphologique de brevets. De multiples interventions humaines sont présentes dans les travaux de Yoon et Kim [YK12]. On peut citer par exemple le processus de prétraitement ainsi que le dépistage de structures sujet-action-objet(SAO). D'autres travaux mettent en avant les interventions d'experts dans le processus d'extractions de connaissances. Par exemple Geum et al. [GJS13] qui décrivent l'extraction de mot-clés où le jugement des experts dans la définition des mots-clés est une étape essentielle, ou encore Lee [LKS+14], qui avec une approche itérative, assistée là aussi par des experts, permet de créer, sélectionner et d'étendre l'ensemble des structures SAO.

Issues du traitement du langage naturel (TAL), les approches automatisées utilisées comprennent des techniques statistiques, linguistiques et sémantiques, telles que les schémas séquentiels généralisés [LAS97], la reconnaissance d'entités nommées [LSLL09, GU10], le marquage de discours [AT10, WQZ+15, NSY16], les structures SAO [GM12, LKS+14] et l'analyse sémantique latente [TSV14]. L'objectif est de baliser et filtrer les caractériques extraites, au moyen

3. Disponible sur : `https://github.com/yooper/stop-words`

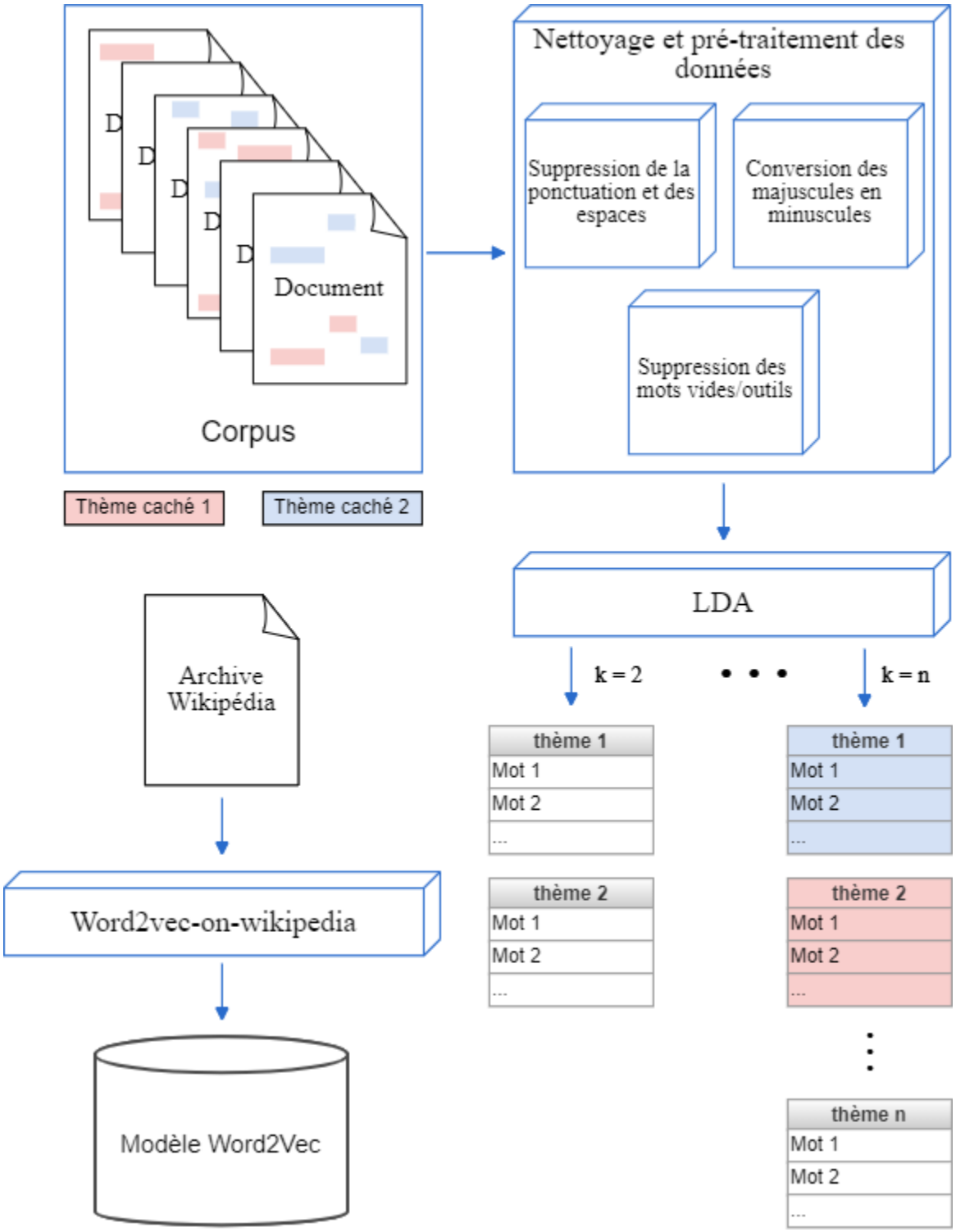


FIGURE *1.2* – ***Représentation des traitements utilisés sur les différentes sources de données. Les résumés des articles référencés par Mühlroth [MG18] correspondent au corpus. Nous appliquons différentes techniques de nettoyage pour rendre les données exploitables par l'approche de modélisation thématique utilisée. Nous appliquons*** LDA, ***nous lançons plusieurs expérimentations en faisant varier le nombre de thèmes. Pour notre approche de plongement de mots qu'est*** Word2Vec, ***nous utilisons un dump des articles provenant du Wikipédia Anglais pour entrainer un modèle au moyen de l'outil Word2vec-on-wikipedia.***

du text mining, pour projeter les données textuelles non structurées dans un modèle d'espace vectoriel structuré afin d'appliquer des techniques d'extraction.

L'utilisation de caractéristiques sur une base entièrement manuelle [Lee08], souvent présentée comme un ensemble de mots-clés déterminés par des experts du domaine, ou le filtrage par un expert humain des mots-clés ou caractéristiques générés automatiquement [AT10, LJP11, YK12] sont des étapes importantes des approches manuelles. Cette participation d'un expert pour la projection et la transformation introduit un biais dans les données, similaire à la création de requêtes lors de la collecte de données [GCPP05, MdFdA+14, HZM+15].

Méta-analyse bibliographique - Mise en œuvre

Pour l'expérimentation proposée, notre étape de projection et transformation de données est composée de deux approches automatisées issues du traitement du langage naturel :

— Modélisation thématique avec *LDA* : cette méthode permet de projeter les données textuelles non structurées. L'idée sous-jacente est que chaque document est modélisé comme un mélange de Thèmes. Nous utilisons différentes valeurs pour le paramètre k définissant le nombre de thèmes *a priori* pour ainsi obtenir plusieurs partitionnements des données (cf. Figure 1.2). Dans notre expérimentation k évolue de 2 à 12.

— Plongement de mots avec *Word2Vec* : dans ce modèle présenté par Mikolov [MSC+13], les mots sont représentés par des vecteurs caractéristiques des relations contextuelles qui les relient entre eux. Nous avons entrainé un modèle sur l'ensemble des documents du Wikipédia Anglais à l'aide de l'outil Word2vec-on-wikipedia [4].

1.2.4.1.4 Méthodes d'exploration de données (*data mining*)

Approche générale - état de l'art

Comme il a été proposé par Mühlroth [MG18], nous présentons les différentes approches de la littérature selon la classification suggérée par Fayyad [FPSS96] présentée en introduction (voir 1.2). Afin de dégager les méthodes et *tendances* méthodologiques utilisées dans différents articles, nous les étudions selon les perspectives combinées des objectifs d'analyses et des méthodes d'exploration de données. D'une manière générale, les processus d'exploration de données peuvent nécessiter de multiples itérations pour obtenir les résultats souhaités en fonction des paramètres de l'algorithme choisi. Ils peuvent également nécessiter des retours en arrière pour ajuster les étapes en amont [FPSS96], et plusieurs méthodes d'exploration de données peuvent être utilisées de manière conjointe.

Dans la littérature, des coocurrences intéressantes de méthodes d'exploration de données, d'objectifs d'analyses et de type de sources de données partagent des modèles communs et des *tendances* méthodologiques similaires.

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

— **Détection d'anomalies**

La détection d'anomalies est largement représentée dans l'extraction des *signaux faibles* et des *tendances* technologiques (voir par exemple les travaux de [SKT+11, WM12, NSY16]). On peut citer les problématiques d'analyse de cooccurence de mots [Lee08], de citations [SKT+11], de co-citations [GALR16]. Elle est utilisée également dans la recherche à partir de séries temporelles (comme par exemple l'analyse de l'évolution de mots [dCdSF06, FC08, AT10, PKB+11, TS12]) ou de détection de points anormaux [TTY14]). La détection d'anomalies, utilisée dans des processus manuels avec participation d'experts, permet la création de feuilles de route technologiques (par exemple dans le cadre de méthodes d'analyse SAO [WQZ+15].

Les publications scientifiques et les données de brevets sont très souvent utilisées comme sources de données, notamment lors de la recherche de *signaux faibles* technologiques. Les informations issues du web ou des réseaux sociaux sont nettement moins employées. Dans le domaine des *tendances* sociales, la détection d'anomalies est très présente durant les années 2010 à 2015 et permet l'analyse de l'évolution des mots [CCS13, PWY+13, KKHR15]. Elle laisse place à d'autres méthodes par la suite.

Quelques travaux utilisent aussi cette approche dans les domaines de la recherche de *tendances* dans les secteurs politiques [RRSZ14] et économiques [PNLM13, KLK+15]

— **Regroupement**

Les techniques de regroupement sont souvent adoptées dans la recherche de *signaux faibles* technologiques lorsque les sources de données sont des publications scientifiques (par exemple K-Means [VBV10], regroupement par vecteur support [JPJ12b], k-médoïds [MHCS15], analyse SAO sémantique [GM12, YK12, LKS+14]).

Cette approche est utilisée dans la recherche des *tendances* sociales sur des sources issues majoritairement du web [CTT06, BXMH15] et des réseaux sociaux [BOMOC14, FZYL14, BXMH15] (par exemple K-Means [BOMOC14], regroupement multirelation [FZYL14] et regroupement multisources [BXMH15]).

Le regroupement est faiblement utilisé dans le recherche de *tendances* technologiques où il est employé avec des sources de données utilisant principalement des mos-clés [WSLS10, JY15, MP15].

Cette approche est également utilisée pour l'étude de *tendances* politiques [MZ05, GV17] et économiques [DCWX10].

— **Classification**

Les approches de classification sont principalement utilisées pour l'analyse et l'extraction de *tendances* sociales (par exemple *LDA* ou de manière plus générale, les approches de modélisation probabiliste [LPZ15, WLT+15, YLLL16, XZJ+16]). Face à un flux constant de données textuelles, la mise à jour régulière du corpus pour détecter les nouvelles *tendances* sociales justifie l'utilisation de modèles probabilistes capables de sauvegarder et conserver leur état du modèle [LCB12, APM+13].

Ces méthodes sont faiblement utilisées pour la modélisation des *tendances* en R&D

[WM06, BEG09] et ne sont pas présentes dans les travaux de recherche de *signaux faibles* technologiques. Dans le domaine des *tendances* technologiques, l'analyse des modèles de convergence s'est principalement appuyée sur des analyses de cooccurence [CL11], comme par exemple la co-classification des codes IPC des brevets [Cav16].

— **Groupement par similitude / régles d'associations**

Les méthodes de groupement par similitude sont principalement utilisées pour la détection des *tendances* technologiques [LKK10, WCL14, MHKB16, TH16]. On peut citer l'analyse de réseaux (par exemple les réseaux de cooccurences [WCK10], les réseaux SAO [CYK+11], les réseaux de co-citation [BAB13]) ainsi que les approches par règles d'association (extraction des patrons/motifs séquentiels [LAS97], extraction de règles d'association [SLH10, JPJ12a]).

Les travaux de recherche plus récents montrent des méthodes d'exploration de données toujours plus complexes. Concernant la détection de *tendances* technologiques, depuis 2014, on peut noter une évolution des approches, allant de la **détection d'anomalies** et de **regroupement** à des algorithmes utilisant majoritairement des méthodes de **groupement par similitude** [HC14, WCL14, CLLH15, MHKB16, TH16].

Le nombre d'applications augmentant, les méthodes de fouille de données/extraction de connaissances et d'apprentissage automatique deviennent plus spécifiques, plus complexes et fortement dépendant du type des sources de données utilisées. Les travaux sur des sources de données telles que les citations dans les articles scientifiques et les brevets ont augmenté dans le domaine des *tendances* technologiques [KGL+14, HC14, HZM+15, NSL16, TH16]. C'est le cas par exemple de la modélisation des dépendances sous forme d'analyse par des réseaux de citations de brevets [KGL+14]). Ils peuvent nécessiter la participation d'experts [BAB13, KGL+14, HZM+15].

Le groupement par similitude n'est utilisé que rarement pour la détection des *signaux faibles* [LSLL09, RTK+16] ainsi que pour la recherche de *tendances* dans les domaines économiques [WRP+13] et sociales [AOGS13, KTKK15].

— **Régression**

L'utilisation des approches de régression est décrite dans quelques articles sur la détection des *tendances* technologiques [TWTDT11, MdFdA+14, CZZL15] et sociales [LCA15].

Avec cette approche, les données de brevets sont très souvent utilisées comme sources d'information [TWTDT11, MdFdA+14, CZZL15]. Ces données sont principalement étudiées selon leurs évolutions dans le temps [TWTDT11, CZZL15, LCA15].

— **Résumé automatique de texte**

La littérature de Mühlroth [MG18] présente 1 article utilisant l'approche de résumé automatique de texte [BI06]. Dans celui-ci, Bun et Ishizuka [BI06] génèrent des résumés sur les changements perçus dans un domaine au moyen d'agents d'information intégrés à un système de suivi de sujets émergents.

Méta-analyse bibliographique - Mise en œuvre - Détection de *signaux faibles* potentiels

L'expérimentation que nous présentons sur les articles étudiés par Mühlroth [MG18] utilise notre méthode combinant les approches d'exploration de données provenant de la **classification** et de la **détection d'anomalies**. Celle-ci sera détaillée dans le chapitre 2.

Dans l'expérimentation sur les résumés des 91 articles référencés par Mühlroth [MG18], nous obtenons 8 thèmes (cf. Tableau 1.4). Pour chacun des documents, nous identifions son thème d'appartenance en déterminant le nombre de mots communs entre chaque thème et le document. Un document peut être associé à un ou plusieurs thèmes si la différence entre coefficients d'appartenances entre le thème principal et les autres thèmes est inférieure à 5%. On remarque que chaque thème, à part les thèmes 3 et 6, relève d'une même proportion de documents.

Parmi les mots-clés obtenus, on remarque un grand nombre de mots-clés génériques, liés à des méthodes d'analyse ou liés à des domaines "techniques/technologiques". Par exemple, des mots associés à différentes techniques, qu'elles soient industrielles, comme "3d" utilisé régulièrement avec "printing" [PKL+16], ou relevant du *data mining* comme "tf", faisant soit référence à "Term Frequency" [AT10, NSY16] ou à "technology forecasting" [JPJ12b], ou bien encore "sao" en rapport avec les méthodes d'analyse sujet-action-objet [CYK+11, YK12]. On trouve aussi des mots décrivant différentes technologies telles que "oled", "window" [HC14], "rfid" [TWTDT11], "5g" [NSL16], "tod" pour "technology opportunity discovery" [KGL+14] régulièrement utilisé dans le domaine industriel.

D'autre mots n'ayant pas de liens apparents avec ces thématiques sont aussi présents dans ses mêmes articles. Ils montrent cependant une certaine proximité avec le vocabulaire technologique comme "pdf" associé à "Proportional Document Frequency" [NSY16], "remarkable" [AT10] et "turning" [CZZL15].

On trouve des mots du domaine de la santé et du médical dans les thèmes 3, 5, 6 et 8, tels que "health-related", "medicine", "medically-related", "health", "medical", "pharma", "biomedical", "pharmaceutical" [SKT+11, LZL+13, PWY+13, PNLM13, WRP+13, CLLH15, MP15, GALR16]. On remarque que certains mots du même thème ne semblent pas reliés à la sémantique médicale. Ils sont présents cependant dans les mêmes documents : "engineering" [CLLH15], "Wikipedia" [PWY+13] et "convergence" [PNLM13, WRP+13].

Cette analyse permet d'identifier des groupes de mots révélant différentes sémantiques caractérisant les thèmes obtenus, bien que ceux-ci ne soient pas parfaits. La complexité d'identification vient de la taille des documents. En effet, les résumés des articles sont courts et décrivent seulement de manière concise leur sujet. Cela implique un faible nombre de mots du domaine de chaque document et donc une difficulté de leur identification.

L'étude de ces premiers résultats expérimentaux (cf. Tableau 1.4) ne permet pas de discerner de *signaux faibles*. Seul le thème 1 présente un fort degré de cohérence d'après l'indicateur construit via l'approche *Word2Vec* (cf. chapitre 2). Il s'avère donc difficile de dégager de cette

première analyse un thème susceptible d'être un *signal faible*. Il ne s'agit pour l'instant que de *signaux faibles* potentiels. Une seconde étape est nécessaire afin de suivre l'évolution de ces thèmes et ainsi mettre en évidence un *signal faible* potentiel devenant *signal fort*.

1.2.4.1.5 Exploration et visualisation des *signaux faibles* détectés et *tendances* émergentes

Approche générale - état de l'art

La visualisation des données consiste à transformer des données complexes en représentation visuelle simple facilitant ainsi leur compréhension et leur exploitation. L'objectif de cette stratégie est double : d'une part aider à construire grâce aux données intermédiaires la suite des traitements, et d'autre part, à guider la prise de décision à partir des données résultats.

En terme de représentation, on retrouve les formes classiques : représentation chronologique, graphiques linéaires, circulaires, en barres, en colonnes, les tableaux, la visualisation de séries chronologiques ou de lignes de temps et les visualisations sous la forme de réseaux ou de graphes.

D'autres techniques de visualisation moins classiques sont utilisées comme les cartes thématiques, les feuilles de route, les nuages, les radars, les cartes du monde ou les arbres de classification. Ces données fournissent des informations contextuelles supplémentaires aux experts du domaine pour identifier des opportunités technologiques.

L'identification des techniques de visualisation, en aggrégeant les approches d'exploration de données et les objectifs d'analyses, n'ont pas permis de révéler de *tendances*. Les formes classiques de représentation sont cependant majoritairement utilisées à la fois pour les données sources et les résultats d'analyses.

Ces données, résultat de l'analyse, se présentent sous la forme de critères quantitatifs (par exemple, les travaux de Nguyen [NSY16] interprètent les poids normalisés d'occurence de termes comme des *tendances* technologiques) ou qualitatifs nécessitant des outils de visualisation spécifiques (cartes de domaine de Lee [Lee08], analyse multidimensionnelle de Huang [HC14], profils de veille technologique de Vengelers [VBV10])

Méta-analyse bibliographique - Mise en œuvre

Dans cette dernière étape du processus d'extraction de connaissance, nous utilisons les données obtenues lors de la première phase de traitement pour alimenter un système d'*agent mining*. L'objectif est d'extraire, grâce à un suivi longitudinal, les *signaux faibles* (cf. chapitre 3).

Des agents de recherche enrichissent la base documentaire de nouveaux documents et mots grâce à des requêtes sur des moteurs de recherches. Dans le cadre de cette expérimentation portant sur les articles relatifs aux *signaux faibles* et *tendances* émergentes référencés par Mühlroth [MG18], nous construisons une base de données composée d'articles extraits de WoS pour la

TABLE *1.4* – ***Détail des thèmes obtenus par notre approche de modélisation thématique multi-niveaux. Le thème 1 semble le plus cohérent d'après l'indicateur construit via l'approche*** Word2Vec***.***

Nom du thème	thème 1				thème 2			
Nombre de documents	12				11			
Cohérence du thème (I_1)	1474				1063			
LDA $k =$	9				8			
	TF-IDF		LDA		TF-IDF		LDA	
Premiers mots du thème	3d	0,23543	technology	0,05275	cnt-fed	0,14741	field	0,02431
	printing	0,21660	3d	0,02867	gatherings	0,10175	analysis	0,02210
	vacant	0,12282	printing	0,02638	fronts	0,09604	fronts	0,01436
	fca	0,09598	paper	0,02179	fca	0,09598	citation	0,01326
	tf	0,08939	technological	0,01950	journal	0,09022	clusters	0,01326
	cgee	0,08619	authors	0,01720	path	0,08239	scientific	0,01215
	km-svc	0,06699	forecasting	0,01720	trajectory	0,08239	study	0,01215
	ma-lda	0,06353	patent	0,01491	window	0,06265	network	0,01215
	cnt	0,05897	technologies	0,01491	oled	0,06265	fields	0,01105
	lattice	0,05759	vacant	0,01261	eighteen	0,06265	studies	0,00994
	analysed	0,05716	future	0,01147	sliding	0,06265	evolution	0,00994
	increases	0,05097	areas	0,01147	conceptual	0,06179	method	0,00994
	gartner	0,05040	trends	0,01147	engineering	0,06151	trends	0,00994
	analyses	0,04543	objective	0,01032	inter-relationships	0,06014	mass	0,00884
	mot	0,04466	data	0,01032	research-front	0,05955	publications	0,00884
	goal	0,04317	tf	0,00917	cnt	0,05897	science	0,00884
	trm	0,04317	map	0,00917	neviewer	0,05858	concepts	0,00884
	alternatives	0,04157	analysis	0,00917	analysed	0,05716	emerging	0,00884
	phases	0,04101	evolution	0,00917	descriptor	0,05435	cluster	0,00773
	paths	0,03869	model	0,00917	profiling	0,05435	subject	0,00773

Suite à la page suivante

TABLE 1.4 – suite de la page précédente

Nom du thème	thème 3				thème 4			
Nombre de documents	7				12			
Cohérence du thème (I_1)	758				646			
LDA $k =$	9				9			
	TF-IDF		LDA		TF-IDF		LDA	
Premiers mots du thème	nutrition	0,12727	convergence	0,06375	weak	0,10250	business	0,024038
	convergence	0,06694	industry	0,05100	rule	0,07167	intelligence	0,024038
	fusion	0,06650	medical	0,01821	event	0,06878	approach	0,020833
	5g	0,05569	industries	0,01821	turning	0,06414	time	0,019231
	acquisitions	0,05245	nutrition	0,01639	anticipative	0,06179	weak	0,016026
	ict	0,05245	technological	0,01457	neviewer	0,05858	trend	0,016026
	mergers	0,05245	ipc	0,01275	vib	0,05674	change	0,016026
	subdomains	0,05126	co-classification	0,01093	advance	0,05569	managers	0,014423
	health	0,04872	phenomenon	0,01093	warning	0,05569	mining	0,014423
	medical	0,04657	fusion	0,01093	consumers	0,05276	activities	0,014423
	analyses	0,04543	based	0,01093	organization	0,04881	trends	0,014423
	higher	0,04256	food	0,00911	stock	0,04799	signals	0,012821
	telecommunications	0,04252	sector	0,00911	signals	0,04720	approaches	0,012821
	pharma	0,04242	promising	0,00911	rules	0,04173	data	0,012821
	biomedical	0,03823	telecommunications	0,00911	change	0,03888	rule	0,012821
	obscure	0,03823	cases	0,00911	descriptions	0,03823	knowledge	0,012821
	drug	0,03823	fields	0,00911	spaces	0,03823	organization	0,009615
	pharmaceutical	0,03743	markets	0,00729	delivery	0,03823	competitive	0,009615
	settings	0,03743	future	0,00729	restricted	0,03823	existing	0,009615
	anticipate	0,03743	impact	0,00729	encompassing	0,03823	event	0,009615

Suite à la page suivante

TABLE 1.4 – suite de la page précédente

Nom du thème	thème 5				thème 6			
Nombre de documents	13				18			
Cohérence du thème (I_1)	533				422			
LDA $k =$	8				11			
	TF-IDF		LDA		TF-IDF		LDA	
Premiers mots du thème	rfid	0,19028	rfid	0,02773	bursty	0,08551	twitter	0,03669
	remarkable	0,09736	frequency	0,02284	outlierness	0,08019	data	0,02935
	tf	0,08939	indices	0,01958	blogs	0,07614	social	0,02516
	broadcasts	0,08007	document	0,01631	eigen-trends	0,06940	tweets	0,02516
	technical	0,07279	technical	0,01631	ma-lda	0,06353	propose	0,01782
	turning	0,06414	promising	0,01305	collective	0,05546	real	0,01468
	phrase	0,06229	process	0,01305	parliamentary	0,05370	network	0,01468
	taxonomy	0,06096	phrase	0,01142	consumers	0,05276	users	0,01258
	analysed	0,05716	china	0,01142	medically-related	0,04957	user	0,01258
	indices	0,05574	terms	0,01142	health	0,04872	topics	0,01258
	5g	0,05569	subjects	0,00979	creation	0,04650	public	0,01048
	library/digital	0,05435	documents	0,00979	social-network	0,04444	techniques	0,01048
	libraries	0,05435	applications	0,00979	anomaly	0,04411	media	0,01048
	disseminated	0,05338	explore	0,00816	topicsketch	0,04276	number	0,01048
	broadcast	0,05338	cluster	0,00816	lda	0,04236	trends	0,01048
	intensity	0,05338	occurrence	0,00816	opinions	0,04196	focus	0,00943
	medicine	0,05276	content	0,00816	web	0,04052	knowledge	0,00943
	consumers	0,05276	years	0,00816	wikipedia	0,04021	posts	0,00839
	themes	0,05097	papers	0,00816	mvtd	0,04010	detect	0,00839
	query	0,04534	collection	0,00816	real-time	0,04004	wikipedia	0,00734

Suite à la page suivante

TABLE 1.4 – suite de la page précédente

Nom du thème	thème 7				thème 8			
Nombre de documents	12				12			
Cohérence du thème (I_1)	405				395			
LDA $k =$	9				11			
	TF-IDF		LDA		TF-IDF		LDA	
Premiers mots du thème	cnt-fed	0,14741	patent	0,08201	bursty	0,08551	topics	0,05587
	roadmap	0,10330	technology	0,05513	generative	0,07276	topic	0,05214
	sao	0,08627	patents	0,03653	modelling-based	0,06634	detection	0,04004
	outlierness	0,08019	technological	0,02136	vocabulary	0,06634	emerging	0,03445
	nest	0,08007	opportunities	0,01654	injected	0,06634	time	0,02421
	nedd	0,07646	analysis	0,01654	mechanism	0,06634	hot	0,02048
	smes	0,07395	development	0,01516	implements	0,06634	online	0,01955
	two-stage	0,07395	identify	0,01447	in-built	0,06634	model	0,01490
	morphology	0,06236	technologies	0,01447	subtopics	0,06634	detect	0,01304
	ma	0,06236	method	0,01309	microblog	0,06397	method	0,01117
	novelty	0,06180	potential	0,01103	ma-lda	0,06353	events	0,01117
	attributed	0,06014	data	0,01103	novelty	0,06180	microblog	0,01024
	tod	0,05692	identification	0,01034	health-related	0,05954	tracking	0,01024
	vacancy	0,05435	citation	0,00965	summary	0,05897	detecting	0,00931
	function	0,05370	innovation	0,00896	lattice	0,05759	experiments	0,00838
	membrane	0,05370	network	0,00896	analysed	0,05716	novelty	0,00745
	pattern	0,05276	proposed	0,00758	elaboration	0,05328	word	0,00745
	tempest	0,05165	extracted	0,00758	flickr	0,05328	based	0,00745
	creation	0,04650	emerging	0,00758	organization	0,04881	news	0,00745
	security	0,04411	existing	0,00689	health	0,04872	features	0,00652

période d'avril 2017 jusqu'à septembre 2020. Ces articles sont placés dans Elasticsearch. Sur la base des thèmes extraits initiaux (période 1997-2017), des agents de recherche forment ensuite des requêtes sur la base de données Elasticsearch pour enrichir le corpus de documents. Nous construisons alors des graphiques et tableaux de résultats pour suivre l'évolution des thèmes.

Plusieurs informations sont obtenues, notamment, pour chaque mot requête, le nombre de documents qu'il rapporte (cf. Tableau 1.5). Comme nous le détaillerons dans le chapitre 3, un ou plusieurs mots de chaque thème sont choisis aléatoirement pour construire les requêtes sur la base définie dans ElasticSearch. Dans le cas de ce suivi historique, l'ensemble des requêtes est construit à partir des 40 premiers mots de chaque thème. Plus le mot d'un thème rapporte de documents, plus ce mot est pertinent, relativement à la base, et participe de surcroit à renforcer ce thème.

Mots	Nb document(s)	Mots	Nb document(s)
rights	61	computing	24
higher	47	software	23
internet	43	rules	23
web	41	matrix	22
health	41	open	22
industry	40	range	20
fusion	40	goal	18
technical	40	generally	18
change	38	security	18
lda	34	outperforms	18
environmental	33	detailed	18
analyses	33	tweets	17
pattern	32	frequent	17
medical	30	samples	16
real-time	29	increases	16
convergence	28	signals	16
function	27	event	15
platform	25	extensive	15
engineering	24	communities	15
point	24	points	15

TABLE *1.5* – ***Nombre de documents obtenus après requêtes à partir de mots-clés dans un moteur de recherche (ElasticSearch) composé des 537 résumés d'articles. Ces derniers couvrent la période 2017-2020 et ont été extraits à partir de la même requête utilisée par Mühlroth [MG18] sur la base de données de WoS (cf. Tableau 1.3). Cette liste comprend les requêtes qui ont permis la récupération d'au moins 15 documents supplémentaires.***

A partir des tableaux 1.5 et 1.6 (cf. Figures 1.3 et 1.4), nous pouvons effectuer les constats suivants (on étudie les 20 premiers mots de chaque thème et on se fixe un seuil arbitraire de documents rapportés à 15) :

— Ceux sont les mots des thèmes 3 et 4 qui rapportent le plus de documents (> 15)

— Parmi les 40 mots rapportant au moins 15 documents, seulement 20 sont présents parmi les premiers mots séléctionnés de chaque thème (cf Figure 1.7).
— les 40 mots de la liste (dont le nombre de documents rapportés est supérieur à 15) rapportent 60% des documents (soit 1076 sur un total de 1807 rapportés). Les 151 autres ne rapportent que 40% de documents (soit 731).

Le tableau 1.7 répertorie les mots rapportant le plus grand nombre de documents :

— Les mots qui relèvent de la sémantique santé tels que "health" (41 documents récupérés), "medical" (30 documents récupérés) ou qui s'y rapporte comme "engineering" (24 documents récupérés) ou "convergence" (28 documents récupérés) rapportent 7% des documents (soit 123 sur un total de 1807 rapportés) (cf. Figure 1.3 et Tableau 1.7). Les documents relatifs à ce domaine sont donc maintenant fortement présents sur la période 2017-2020 alors qu'ils ne l'étaient pas auparavant. Ce domaine d'étude non référencé comme tel dans l'article de Mühlroth [MG18] représente donc sur la période 2017-2020 un attrait fort dans la littérature. On peut raisonnablement considérer que les mots-clés de ce domaine obtenus à partir de l'analyse initiale sur la période 1997-2017 sont associés à un *signal faible* potentiel non détecté cependant sur cette période. Ils sont cependant détectés comme *signal faible* sur la période suivante (voir ci-dessous) sur cette période puisque ceux-ci confirment un *signal faible* potentiel sur la période suivante.
— Les mots, dont la sémantique se rapporte aux technologies, récupèrent la majorité des documents. Ceci s'explique par la nature des articles qui relèvent de la détection de *signaux faibles*, des *tendances* émergentes et des méthodes de fouille de données/extraction de connaissances. Mühlroth [MG18] dans ses travaux a montré que la recherche sur les *signaux faibles* et *tendances* émergentes dans le domaine des technologies représente la majorité des articles (58 sur les 91 articles) dont est composée la base. Certains mots de la requête utilisée pour récupérer les articles sont également proches de ce domaine (cf. Tableau 1.3). Nous considérons la sémantique associée, comme un *signal fort*, ce domaine ayant déjà été relevé par Mühlroth sur la période 1997-2017.

	thème 1	**thème 2**	**thème 3**	**thème 4**	
Itération 0	12	11	7	12	
Itération 1	104	81	170	150	

	thème 5	**thème 6**	**thème 7**	**thème 8**
	13	18	12	12
	82	112	88	72

TABLE *1.6* – ***Nombre de documents attribués à chaque thème. L'itération 0 correspond au document de l'étape de collecte de données du processus d'extraction de connaissances. L'itération 1 rajoute les documents obtenus après les requêtes effectuées par mots-clés dans un moteur de recherche (ElasticSearch).***

A partir du tableau 1.8 présentant certains thèmes obtenus après une nouvelle itération de la modélisation thématique multi-niveaux sur le nouveaux corpus de documents (corpus initial et documents rapportés), nous remarquons des changements dans les thèmes :

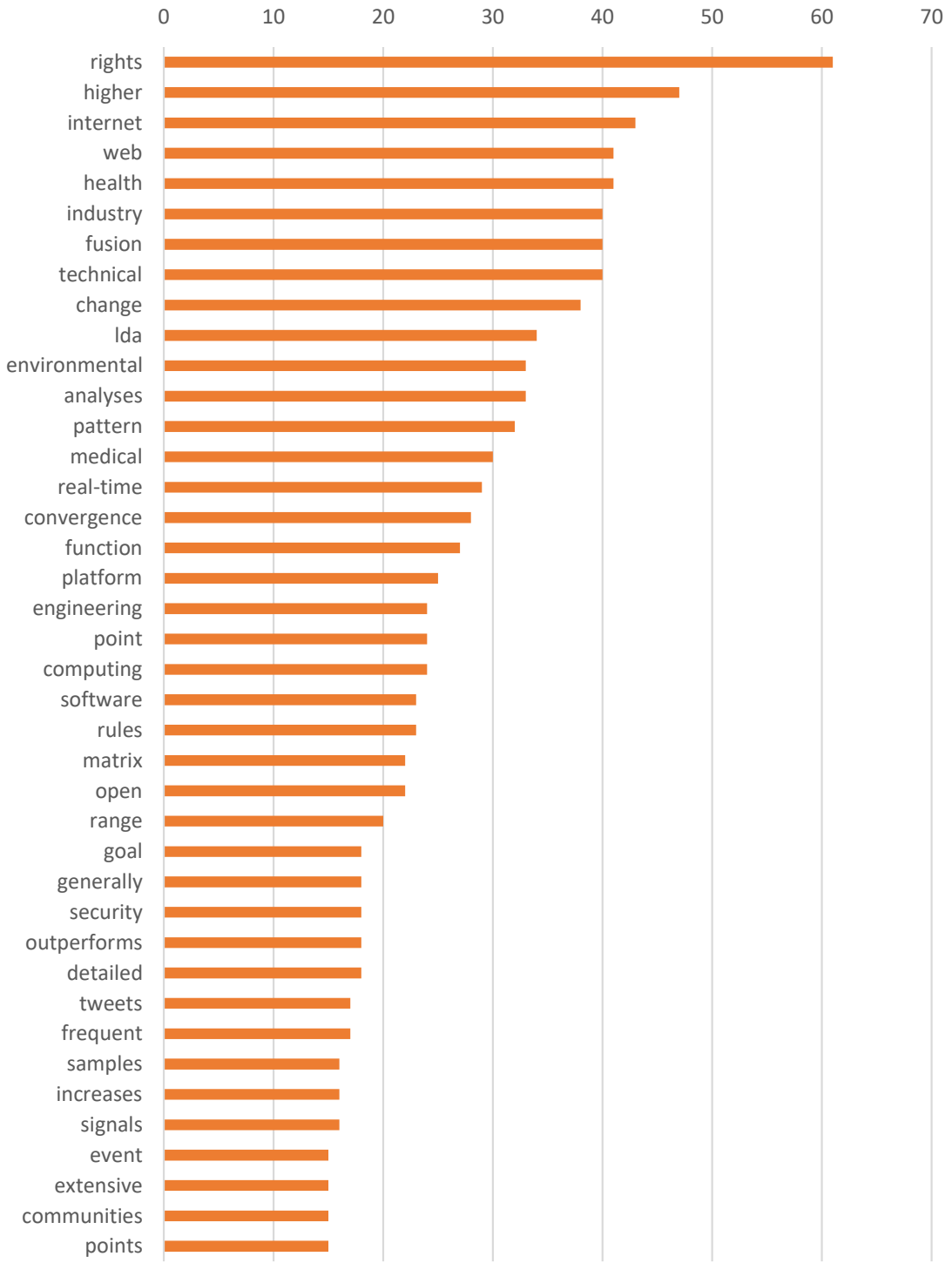


FIGURE *1.3* – ***Représentation graphique du nombre de documents obtenu après requêtes des mots dans un moteur de recherche***

— Les mots qui relèvent du domaine de la santé ont fait émerger un thème portant uniquement sur ce domaine avec des mots tels que "cancer", "disease", "gene", "treament", "drug", "patient", "biological" en plus des mots présents dans l'itération précédente. Ce domaine d'étude non référencé comme tel dans l'article de Mühlroth [MG18] mais identifié sur la période 1997-2017 dans la méta-analyse bibliographique s'avère donc confirmé

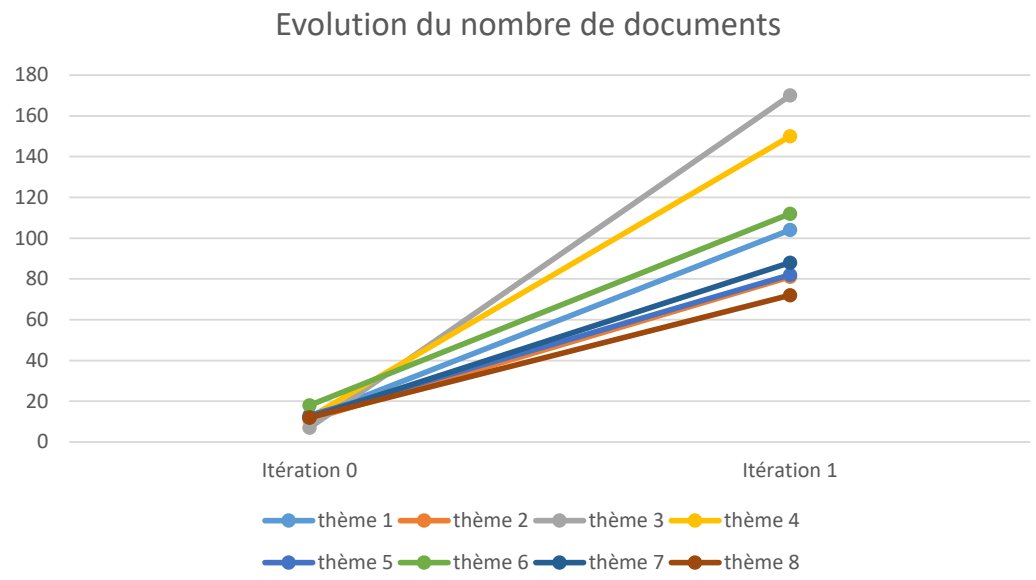


FIGURE *1.4* – ***Représentation graphique du nombre de documents attribués pour chaque thème (cf. 1.6). L'itération 0 correspond au document de l'étape de collecte de données du processus d'extraction de connaissances. L'itération 1 rajoute les documents obtenus après les requêtes effectuées par mots-clés dans un moteur de recherche (ElasticSearch). Les taux de croissance sont respectivement*** 8.7 ***(thème 1),*** 7.4 ***(thème 2),*** 24.3 ***(thème 3),*** 12.5 ***(thème 4),*** 6.3 ***(thème 5),*** 6.2 ***(thème 6),*** 7.3. ***(thème 7) et*** 6.0 ***(thème 8).***

sur la période 2017-2020 comme étant un *signal faible* potentiel. Ce domaine est en forte augmentation depuis les récentes crises sanitaires tels que la pandémie de Covid-19 et l'épidémie de maladie à virus Ebola, ce qui suggère qu'il s'agissait en fait du *signal faible* sur cette période 2017-2020.

— Des nouveaux thèmes portant sur deux domaines différents mais proches sont apparus. Le premier (thème 1) porte sur les études des éventuels impacts des changements climatiques et météorologiques sur l'environnement. Des mots comme "climate", "rainfall", "stations", "spatial", "précipitation", "river", "sea" sont associés à d'autres comme "series", "change", "study", "variability", "detection". Les mots se rapportent à l'étude des changements climatiques et météorologiques. Ces sujets font partie des enjeux sociétaux mais aussi scientifiques d'aujourd'hui : étude des conditions météorologiques et des effets sur la production agricole et alimentaire, élévation du niveau des mers, risques d'inondations et des conséquences à l'échelle mondiale. Le second thème se rapporte aux matériaux et aux impacts environnementaux engendrés par leur utilisation. Des mots dont la sémantique se rapporte aux matériaux comme "sédiment", "pb" pour le plomb, "hg" pour le mercure, "zn" pour le zinc, "minéral", "cu" pour le cuivre, "cd" pour le cadmium et "ore" sont présents avec "mining", "pollution", "waste", "air", "transport". Nous considérons la sémantique associée, comme un *signal faible* potentiel. Dans l'itération précédente, le seul mot parmi les thèmes qui se rapportent à ce domaine était "environmental".

L'utilisation de notre chaine de traitement, dans le but d'étudier plus en profondeur les

	thème 1	thème 2	thème 3
20 premiers mots triés par valeur de *tf-idf*	3d	cnt-fed	nutrition
	printing	gatherings	convergence
	vacant	fronts	fusion
	fca	fca	5g
	tf	journal	acquisitions
	cgee	path	ict
	km-svc	trajectory	mergers
	ma-lda	window	subdomains
	cnt	oled	health
	lattice	eighteen	medical
	analysed	sliding	analyses
	increases	conceptual	higher
	gartner	engineering	telecommunications
	analyses	inter-relationships	pharma
	mot	research-front	biomedical
	goal	cnt	obscure
	trm	neviewer	drug
	alternatives	analysed	pharmaceutical
	phases	descriptor	settings
	paths	profiling	anticipate

thème 4	thème 5	thème 6	thème 7	thème 8
weak	rfid	bursty	cnt-fed	bursty
rule	remarkable	outlierness	roadmap	generative
event	tf	blogs	sao	modelling-based
turning	broadcasts	eigen-trends	outlierness	vocabulary
anticipative	technical	ma-lda	nest	injected
neviewer	turning	collective	nedd	mechanism
vib	phrase	parliamentary	smes	implements
advance	taxonomy	consumers	two-stage	in-built
warning	analysed	medically-related	morphology	subtopics
consumers	indices	health	ma	microblog
organization	5g	creation	novelty	ma-lda
stock	library/digital	social-network	attributed	novelty
signals	libraries	anomaly	tod	health-related
rules	disseminated	topicsketch	vacancy	summary
change	broadcast	lda	function	lattice
descriptions	intensity	opinions	membrane	analysed
spaces	medicine	web	pattern	elaboration
delivery	consumers	wikipedia	tempest	flickr
restricted	themes	mvtd	creation	organization
encompassing	query	real-time	security	health

TABLE *1.7* – ***Liste des thèmes et des mots-clés associés obtenus par modélisation thématique multi-niveaux combinant l'application conjointe d'une modélisation thématique et d'un plongement lexical guidée par une mesure de cohérence. Les mots de couleur verte ont permis la récupération d'au moins 15 documents supplémentaires.***

TABLE *1.8* – ***Détail des thèmes 1, 3 et 4 obtenus par notre approche de modélisation thématique multi-niveaux après une nouvelle itération sur le nouveau corpus de documents. Le thème 1 porte sur le domaine de l'environnement climatique et météorologique, le thème 3 sur les matériaux et leurs impacts environnementaux et le thème 4 sur le domaine de la santé***

Nom du thème	thème 1 / 11				thème 3 / 11				thème 4 / 11			
Nombre de documents	104				170				150			
Cohérence du thème (I_1)	715				594				491			
LDA $k =$	11				9				11			
	TF-IDF		LDA		TF-IDF		LDA		TF-IDF		LDA	
Premiers mots du thème	pan	0,30237	trend	0,03335	rcs	0,26354	mining	0,01199	uc	0,28854	medical	0,01766
	shoreline	0,23064	trends	0,02058	lib	0,20211	concentrations	0,01114	fusions	0,26393	cancer	0,01440
	ffco2	0,19888	series	0,01916	eutrophication	0,19162	sediment	0,01057	rcc	0,25715	clinical	0,01087
	mrp	0,18029	time	0,01703	factory	0,18392	pb	0,00742	dtcs	0,21696	disease	0,01005
	bloom	0,17595	change	0,01348	galena	0,18327	pollution	0,00714	dormancy	0,21696	health	0,00951
	bookmarking	0,14884	climate	0,01064	garbage	0,17778	metal	0,00685	sevoflurane	0,21334	genes	0,00951
	emd	0,14815	rainfall	0,01017	geoai	0,17205	elements	0,00657	procedural	0,20165	treatment	0,00870
	mevd	0,14223	study	0,01017	indigenous	0,15422	hg	0,00628	splicing	0,18286	gene	0,00734
	pheno	0,13334	annual	0,00922	platy	0,14097	zn	0,00628	viruses	0,17978	diseases	0,00734
	irrigated	0,12611	stations	0,00922	e-waste	0,13062	mineral	0,00600	pcos	0,17945	drug	0,00543
	tstm	0,12191	data	0,00922	tgd	0,12750	waste	0,00571	otc	0,17824	patients	0,00489
	hurricane	0,11607	level	0,00828	flux	0,12750	trend	0,00571	cancellation	0,16942	biological	0,00489
	jpcap	0,11511	variability	0,00757	poyang	0,12750	air	0,00514	co-fuse	0,16898	drugs	0,00462
	broadcasts	0,11361	years	0,00710	contact	0,11997	cu	0,00514	deqi	0,16843	chemical	0,00462
	erosivity	0,11032	spatial	0,00686	oil	0,11878	dust	0,00514	ginseng	0,16788	patient	0,00435
	wpvc	0,10922	detection	0,00662	jpcap	0,11511	cd	0,00514	ihr	0,16780	database	0,00435
	tidal	0,10579	period	0,00615	mg/kg	0,11328	observed	0,00514	szgenes	0,16116	cell	0,00435
	eto	0,10407	precipitation	0,00591	wa	0,11295	transport	0,00514	bone	0,15385	study	0,00435
	sub-areas	0,10323	river	0,00568	crystals	0,11278	ore	0,00485	smoking	0,15069	diagnosis	0,00435
	evaporation	0,10159	sea	0,00568	container	0,11112	levels	0,00485	microbiome	0,14869	pubmed	0,00408

articles relevés par Mülroth, a permis de révéler des *signaux faibles* potentiels non référencés comme tel dans ses travaux. Le domaine du médical et de la santé détecté sur la période 1997-2017 est confirmé comme un *signal faible* potentiel sur la période 2017-2020. Deux nouveaux domaines non détectés sur la période 1997-2017 sont apparus durant la période 2017-2020 comme *signaux faibles* potentiels. Il s'agit pour le premier du domaine environnement/climatologie/-météorologie et pour le second du domaine des matériaux et leurs impacts environnementaux (cf. Figure 1.5).

1.2.5 Evaluation des techniques

L'efficacité et l'efficience des algorithmes développés et des solutions proposées doivent être évaluées. Plusieurs solutions ont été mises en œuvre dans les articles :

— **Etude de cas**. Des études de cas sont utilisées comme formes d'évaluations qualitatives telles que des ensembles de données synthétiques [CTT06, TTY14, BXMH15] ou des ensembles de données réelles [TV13, RRSZ14, WQZ+15] ;
— **L'évaluation par des experts externes du domaine**. Les experts évaluent l'efficacité des approches d'exploration de données et fournissent un retour d'information sur la capacité des algorithmes à atteindre l'objectif [VBV10, HC14, TSV14] ;
— **L'utilisation d'indicateurs de performance**. la précision, le rappel et la F-mesure reviennent souvent pour déterminer l'efficacité des approches [LCB12, TSV14, TH16], l'exactitude [KHJJ12, GJS13, KKHR15] ainsi que la perplexité [WLT+15].

Les évaluations des méthodes peuvent entrainer un risque de biais de publication par des mesures quantitatives utilisées délibérément et donnant des valeurs élevées. La performance des approches proposées nécessite une étude empirique détaillée sur l'efficacité des méthodes appliquées sur la base de preuves quantitatives.

1.3 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 [KL17, TV13, APLB05]. La définition que nous adoptons, mise en œuvre par l'utilisation conjointe des approches de "topic modeling" et "word embedding", aussi appelée modélisation thématique multi-niveaux permet, dans l'arborescence des 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 Thème fortement cohérent (unitaire, nouveauté) et disjoint des autres (donc non relié

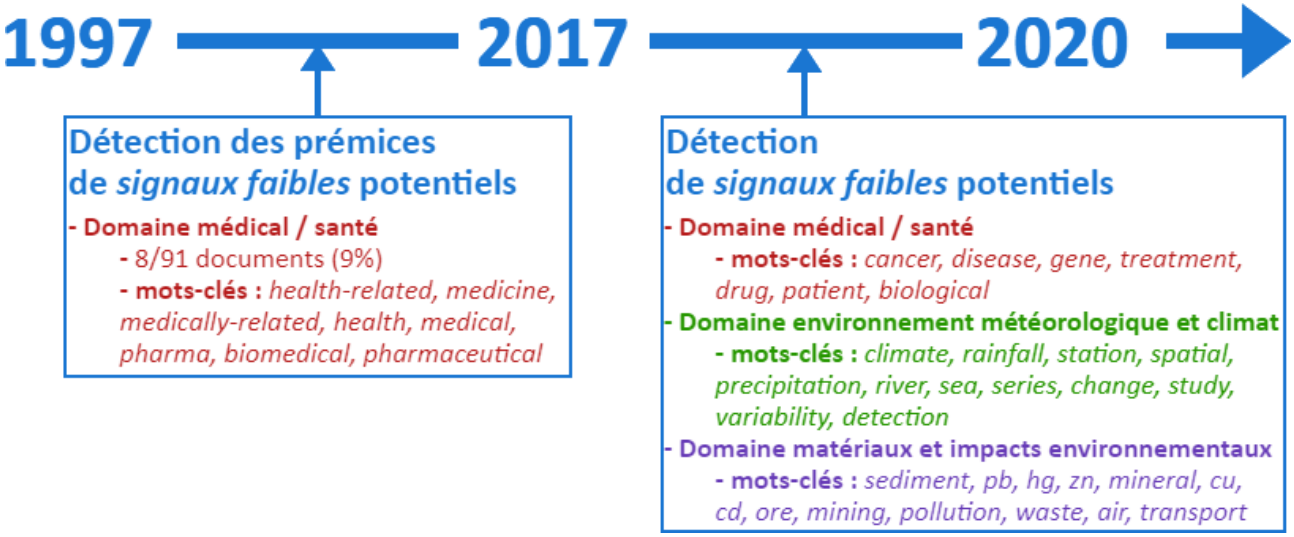


FIGURE *1.5* – ***Frise chronologique de la méta-analyse bibliographique sur les articles de Mühlroth. Sur la période 1997-2017, la chaine de traitement permet de détecter un domaine médical et santé. Ce signal faible potentiel est confirmé après la récupération de nouveaux documents lors de l'étape d'***agent mining***. Deux autres*** signaux faibles ***potentiels sont aussi détectés portant, pour le premier sur le domaine environnement/climatologie/météorologie et pour le second sur celui des matériaux et leurs impacts environnementaux.***

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). Il nous faut mettre en œuvre ensuite une solution d'"agent mining", combinaison des outils de "*data mining*" et "systèmes multi-agents", pour suivre l'évolution de ces mots-clés dans le temps et connecter les thèmes trouvés à un contexte d'information plus large. Les résultats sont présentés sous la forme d'une visualisation interactive et de tableaux montrant l'évolution des *signaux faibles* dans le temps.

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, 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 chaque mot est attribuable en termes de probabilité 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. Cette approche doit prendre en compte l'évolution dans le temps par l'enrichissement de l'information au moyen

de sources externes en ligne, ceci afin de confirmer l'avènement du *signal faible* en un fait observable.

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 de l'étude, de découvrir les mots-clés relevant d'un Thème *signal faible* éventuellement présent.

La suite du traitement est tout aussi complexe. Elle doit permettre de mieux cibler la recherche de nouveaux documents au moyen des mots-clés du thème *signal faible* potentiel. A la difficulté de la recherche de nouveaux documents s'ajoute celle de fournir des indicateurs clairs, sous la forme de visualisation dynamique et de tableaux, permettant à l'utilisateur de prendre des décisions. Le suivi des thèmes potentiellement classés comme *signaux faibles* doit permettre de :

— d'évaluer si un *signal faible* potentiel devient un *signal fort*
— d'enrichir par l'apport de nouveaux documents, l'information existante
— de voir émerger de nouveaux Thèmes

Nous nous focalisons dans le chapitre 2 sur le problème de la modélisation thématique multi-niveaux, première partie de notre solution. Puis dans le chapitre 3, nous présentons notre approche *agent mining*, combinaison d'algorithmes issus du *data mining* et des systèmes multi-agents, construite sur un schéma d'attraction/répulsion. Enfin nous détaillons les résultats obtenus sur plusieurs jeux de données tels que le corpus des projets H2020.

Dans le chapitre 2, 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 soient 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'appuierait fortement sur une chronologie éventuelle des documents. Nous utilisons le corpus comme base initiale pour démarrer nos investigations. La première étape de notre processus combinant classification et plongement lexicale permet d'obtenir les thèmes pertinents du corpus selon notre définition du *signal faible* (cf. figure 1.6). Ces résultats, ainsi que les documents, servent ensuite d'informations d'entrées pour la deuxième partie de la chaine de traitement.

Présenté dans le chapitre 3, l'approche *agent mining* utilise des agents de recherche qui requêtent le web avec les mots-clés des thèmes obtenus (cf. figure 1.7). Ces nouveaux documents récupérés par les agents sont horodatés à $t+1$ par rapport au corpus initial. Cette analyse est

relancée régulièrement afin de générer de nouveaux thèmes sur la nouvelle base documentaire ainsi complétée. Les agents de recherche utilisent les nouveaux thèmes obtenus pour relancer des recherches et ainsi de suite. Ces étapes, alternance de modélisation thématique et de recherche de nouveaux contenus, permettent le suivi des thèmes et la découverte d'autres.

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 ressérrés. 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'information. 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 ressérré (i.e. ne correspondant pas à un pattern textuel). Ceci rejoint les travaux de Ah-Pine [APLB05] qui décrit notamment une information nouvelle comme étant relativement orthogonale aux autres informations contenues dans le corpus.

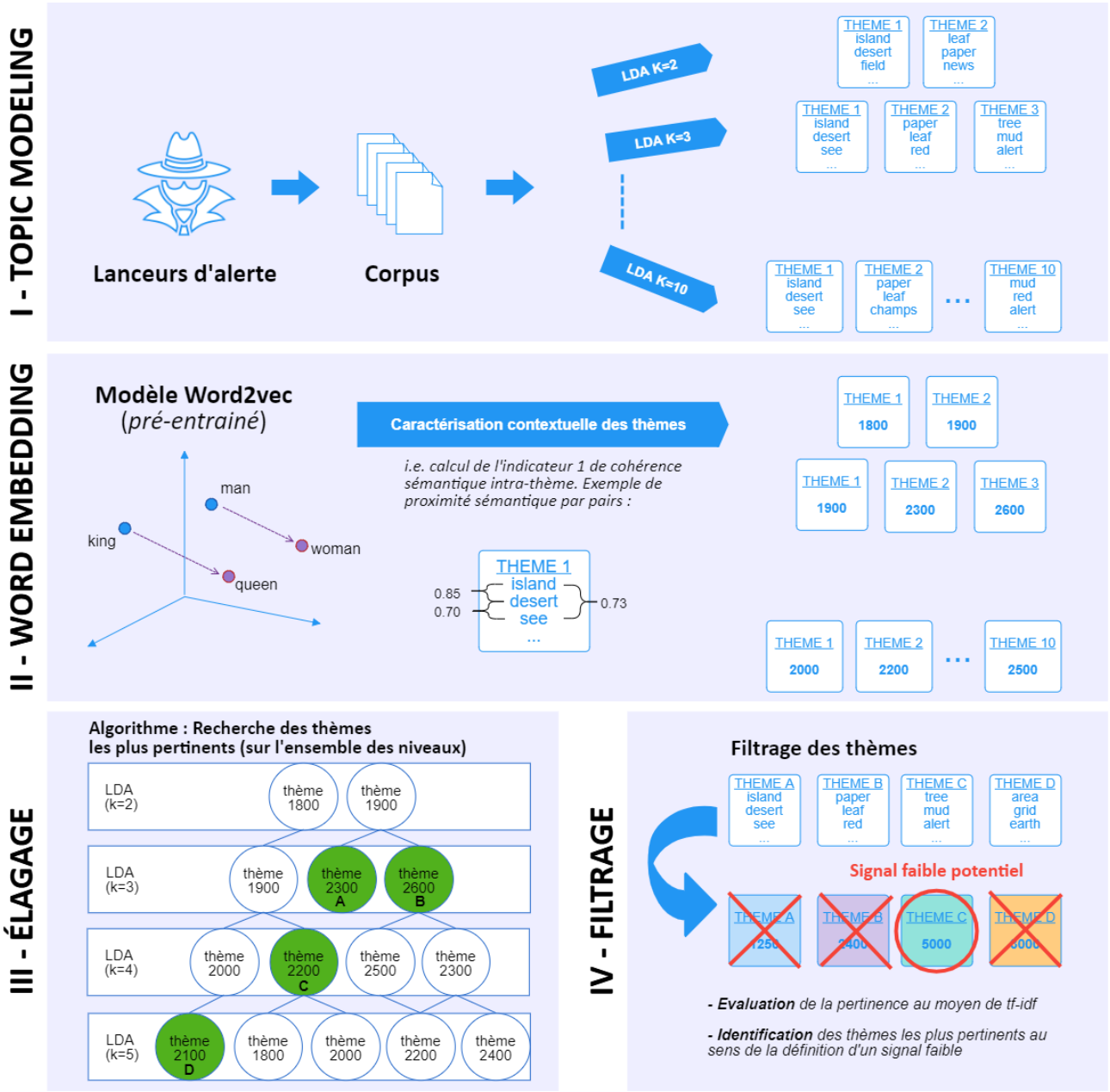


FIGURE *1.6* – ***La première partie de la chaine de traitement utilise une approche de*** Topic Modeling ***et de*** Word Embedding. ***L'algorithme que nous proposons dans le chapitre 2 permet d'extraire les thèmes les plus pertinents sur plusieurs niveaux de*** LDA ***au moyen d'un indicateur basé sur*** Word2Vec. ***Le filtrage permet de récupérer les mots pertinents dans chaque thème selon notre définition du*** signal faible.

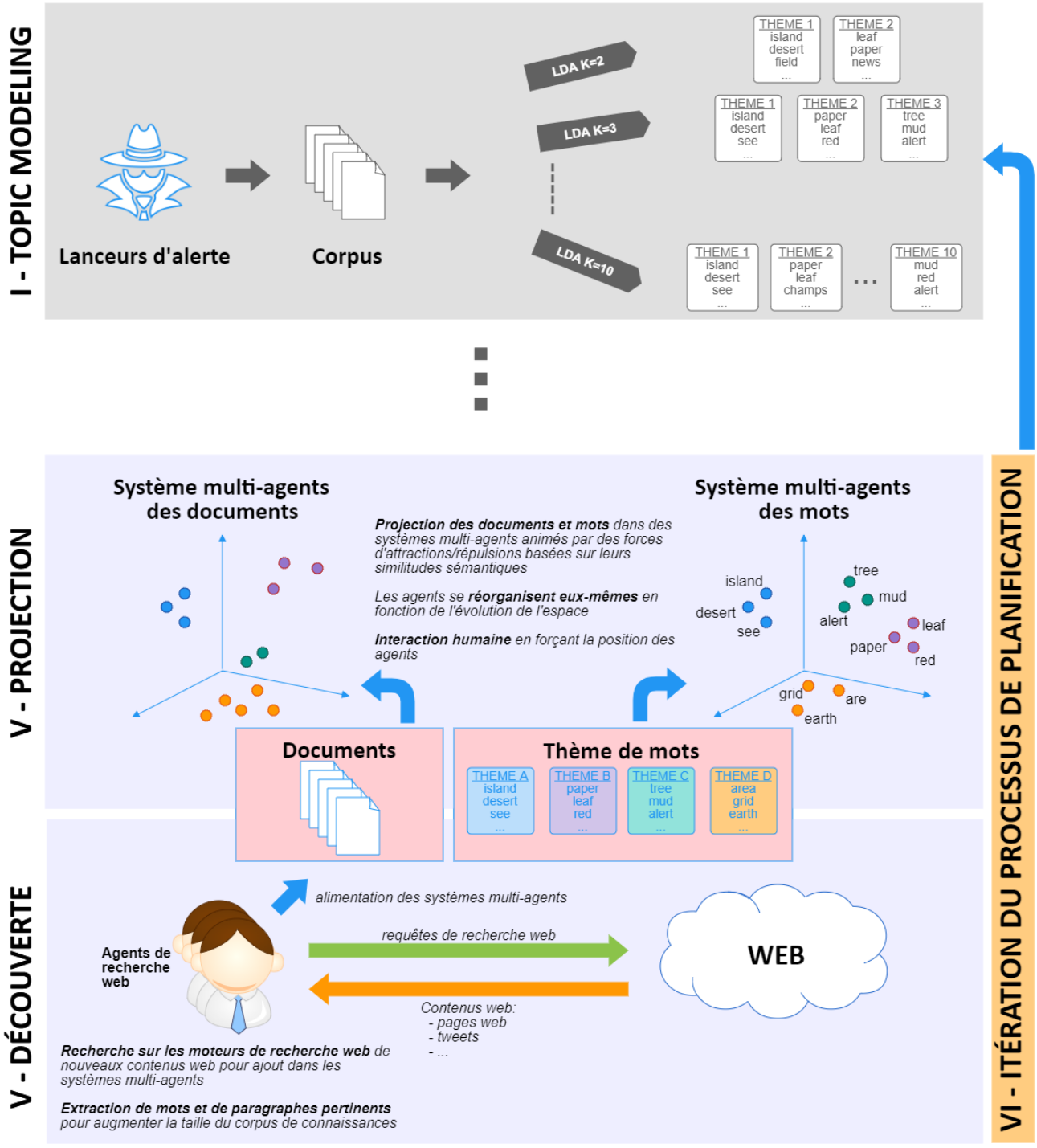


FIGURE 1.7 – ***La seconde partie de la chaine de traitement met en œuvre une solution d'***agent mining***, une combinaison de méthodes de*** data mining ***et de système multi-agents. Les documents du corpus sont projetés dans un système multi-agents animés par des forces d'attraction/répulsion construites sur des similitudes sémantiques (pour les documents) ou de contexte (pour les mots). Des agents de recherche enrichissent le système de nouveaux documents par des requêtes sur des moteurs de recherches (tel que Qwant) à partir des mots-clés des thèmes identifiés.***

Chapitre 2

Modélisation thématique, plongement de mots et exploration d'une collection de documents

Dans ce chapitre, nous présentons la première partie de la solution de la chaine de traitement de données que nous proposons. L'objectif est de répondre au problème de la recherche de *signaux faibles* potentiels dans un corpus de documents représenté sous la forme de sacs de mots. Notre étude de la littérature sur le domaine de la détection de *signaux faibles* ainsi que les méthodes uilisées pour répondre aux différents défis scientifiques de celui-ci nous conduisent à proposer une nouvelle approche, appelée modélisation thématique multi-niveaux, combinant "modélisation thématique" et "plongement de mots" souvent utilisés indépendamment en *traitement automatique des langues* (*TAL*). Nous justifions notamment à partir de l'état de l'art cette approche conjointe afin de répondre à la définition proposée d'un *signal faible*. Elles découlent de paradigmes différents mais s'avèrent complémentaires dans la solution que nous proposons. Nous présentons notre contribution ainsi que des expérimentations sur des corpus artificiels générés et deux corpus de documents réels se présentant respectivement sous la forme d'une base de données Wikipédia et d'une base de données de comptes rendus médicaux. Dans ces travaux, nous utilisons le mot "thème" pour définir les résultats de *LDA* et le mot "Thème" pour définir les ensembles de mots utilisés comme vérité terrain.

Sommaire

1.1 Introduction 19
1.2 Angles d'analyse utilisés 20
1.2.1 Analyse selon les objectifs de détection 21
1.2.2 Analyse selon la source de données 28
1.2.3 Analyse selon les domaines d'analyse 29
1.2.4 Techniques de fouille de données/extraction de connaissances 33
1.2.5 Evaluation des techniques 55
1.3 Positionnement 55

2.1 Introduction

La nécessité grandissante du traitement rapide de l'information conduit au développement de nombreux algorithmes issus de cadres théoriques différents. Des méthodes de traitement et d'extraction efficaces, comme *LDA* ou dérivées du "word embedding" sont régulièrement utilisées.

Le problème adressé dans cette étude est celui de l'évaluation de l'efficacité du modèle *LDA* qui catégorise les documents en un nombre de thèmes défini *a priori*. L'approche que nous proposons pour limiter cet *a priori* consiste à (1) faire varier ce paramètre, (2) estimer le niveau de partitionnement le plus pertinent, et (3) estimer dans l'arbre de profondeur la meilleure collection de mots représentative d'un thème. Dans ce but, nous nous appuyons sur une méthode récente d'apprentissage automatique issue du deep learning de type "word embedding", *Word2Vec*. Elle permet de représenter un mot par un vecteur pour une analyse sémantique. Ainsi deux mots dans des contextes similaires ont des vecteurs proches. Cette approche s'avère donc complémentaire des méthodes s'appuyant sur le comptage de mots dans un document. Elle projette les mots dans un espace vectoriel en fonction du contexte local d'une phrase, au contraire du modèle *LDA* qui trie les mots en fonction de leurs probabilités d'appartenance aux thèmes. Cette complémentarité s'apparente à celle adoptée dans le contexte de la segmentation d'images : utilisation conjointe d'une approche globale par histogramme et d'une approche contextuelle s'appuyant sur les propriétés de voisinage des pixels.

Nous commençons d'abord par décrire le fonctionnement de *LDA* ainsi que sa mise en œuvre. Nous présentons ensuite l'approche *Word2Vec* et ce qu'elle apporte dans l'optimisation de *LDA*. Nous terminons par une présentation de deux algorithmes utilisant l'approche conjointe et une discussion sur les résultats.

2.2 Modèle thématique

Pour gérer l'explosion de la masse d'information des documents électroniques, il faut utiliser de nouvelles techniques ou outils qui traitent, organisent, recherchent, indexent et parcourent de grandes collections de données. De nombreuses techniques automatiques ont été mises au point pour visualiser, analyser et résumer ces derniers [NAXC08].

En s'appuyant sur l'apprentissage automatique et les statistiques, des approches de type "modèle thématique" ont été développées. Ces modèles reflètent les sujets sous-jacents, qui réunissent les documents [AA15]. Les modèles thématiques peuvent être adoptés pour analyser d'autres sources de données que des mots tels que des images, des données biologiques et des données d'enquêtes [BL06]. Pour l'analyse et l'extraction de texte, les modèles thématiques se fondent sur l'hypothèse de sac de mots (i.e. l'information sur l'ordre des mots est ignorée).

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 [DDH90, Hof01, AA15].

2.2.1 L'analyse sémantique latente

L'analyse sémantique latente (*LSA*) est une méthode appartenant au domaine du traitement automatique du language naturel (*Natural Language Processing* ou *NLP*). L'objectif principal de *LSA* est de créer une représentation vectorielle des textes à partir du contenu sémantique. On calcule la similitude entre les textes pour choisir les associations de mots pertinentes. Par le passé, *LSA* portait le nom d'indexation sémantique latente (*Latent Semantic Indexation* ou *LSI*) et s'orientait principalement sur des tâches de récupération d'informations. *LSA* utilise la décomposition en valeurs singulières (*Singular Value Decomposition* ou *SVD*) permettant, outre la réduction de la dimension, de traduire les vecteurs mots et documents dans l'espace des concepts (les concepts sont supposés être orthogonaux). Il est ainsi possible de relier les documents entre eux [DDH90].

Les étapes essentielles de *LSA* sont :

1. la collecte d'un grand corpus de texte pertinent et la division en plusieurs documents ;
2. la création d'une matrice de co-occurrence pour les termes et les documents ainsi que la description des cellules pour le document x et le terme y (où x et y appartiennent respectivement à l'ensemble des documents N et des mots M) ;
3. le calcul de chaque cellule, par une décomposition en valeur singulière sur A, qui donne deux matrices orthonormales U et V et une matrice diagonale Σ.

4. enfin la création des matrices par la sélection des k plus grandes valeurs singulières, ainsi que les vecteurs singuliers correspondants dans U et V. On obtient une approximation du rang k de la matrice des occurences A que l'on appelle A_k (cf. Figure 2.1).

Cette approximation traduit les vecteurs "mots" et "documents" dans l'espace des "concepts".

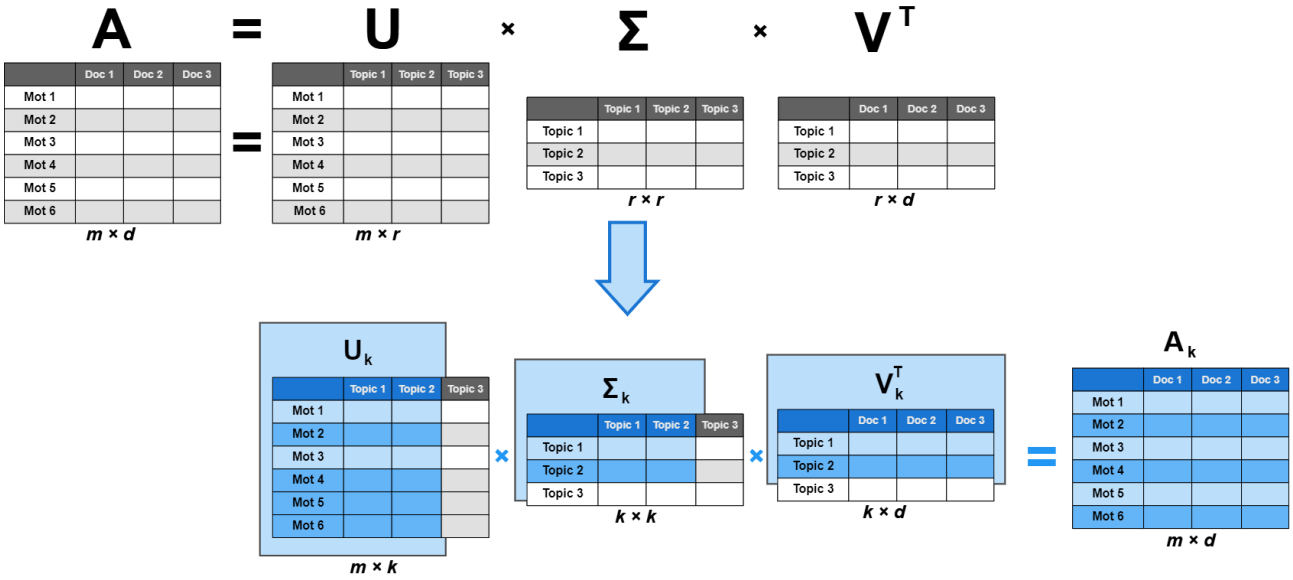


FIGURE *2.1* – ***Représentation graphique de*** LSA

2.2.2 L'analyse sémantique latente probabiliste

L'analyse sémantique latente probabiliste (*PLSA*) se veut être une amélioration de la méthode *LSA*. Jan Puzicha et Thomas Hofmann l'ont introduit durant l'année 1999 [Hof01]. *PLSA* cherche à automatiser l'indexation de documents et repose sur un modèle de classe latente statistique pour l'analyse factorielle du comptage de mots. Cette méthode tente également d'améliorer *LSA* d'un point de vue probabiliste en utilisant un modèle génératif.
L'objectif principal de *PLSA* est d'identifier et de faire une distinction entre les différents contextes d'utilisation des mots sans recourir à un dictionnaire ou un thésaurus. Cela permet de résoudre les ambiguïtés de polysémie, c'est-à-dire les mots à sens multiples, et de révéler des similitudes typiques en regroupant des mots qui ont partagé un contexte commun [Hof01].
PLSA a été utilisé dans de nombreuses applications du monde réel, y compris la vision par ordinateur, et les systèmes de recommandation. *PLSA* a également des applications dans des domaines tels que la récupération d'information et le filtrage, le traitement du langage naturel et l'apprentissage machine à partir de texte. On peut citer :

— la récupération d'images : le modèle PLSA possède des caractéristiques visuelles qu'il utilise pour représenter chaque image en tant que collection de mots visuels à partir d'un vocabulaire visuel discret et fini [RHL09].
— la recommandation automatique de questions [WWC08].

Cependant, la phase d'identification des paramètres augmentant 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 [AA15, KMST08] et la distribution des thèmes reste concentrée sur un nombre limité de sujets [Hof01].

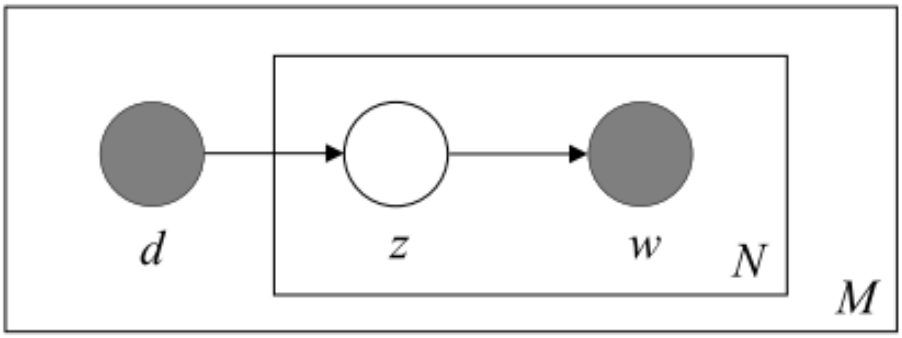


FIGURE *2.2* – ***Représentation graphique de*** pLSA

PLSA est construit sur un mélange de décompositions issues de l'analyse des classes latentes où une classe non observée $z \in Z = \{z_1, \ldots, z_K\}$ est associé à des données co-occurentes comme la distribution des mots $w \in W = \{w_1, \ldots, w_N\}$ au sein d'un document $d \in D = \{d_1, \ldots, d_M\}$ (cf. Figure 2.2).

2.2.3 L'allocation de Dirichlet latente

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" [BNJ03].

Il existe un grand nombre de corpus de documents électroniques, issus du web, de blogs, et de la littérature scientifique, qui posent, en termes d'exploration et d'extraction d'information pertinente, de nouveaux défis aux chercheurs de la communauté *data mining*. De nombreuses techniques automatiques ont ainsi été développées pour visualiser, analyser et résumer ces collections de documents [NAXC08].

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 [AA15, KMST08]. 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* [SSS08]. 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 [RCY06].

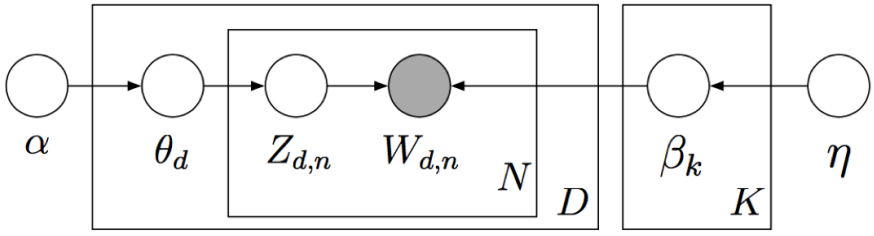


FIGURE *2.3* – ***Représentation graphique de*** LDA

LDA utilise l'approche "Bag of Word" qui traite chaque document d du corpus D défini par $(\mathbf{w_1}, ..., \mathbf{w_D})$ comme un N-uplet de mots, $\mathbf{w_{d,n}} = (w_{d,1}, ..., w_{d,n})$. A chaque mot $w_{d,n}$ est alors associé un thème représenté par la variable $z_{d,n}$ de l'ensemble N du vocabulaire. θ_d représente la distribution de thèmes du document d. Des hyperparamètres, α et η, définissent l'*a priori* sur θ et β où β_K décrit la distribution du thème K parmi les K thèmes (cf. Figure 2.3).
Nous pouvons citer également plusieurs applications s'appuyant sur *LDA*, tels que :

- la découverte de rôles : l'analyse des réseaux sociaux (SCN) est l'étude de modèles mathématiques pour les interactions entre les personnes, les organisations et les groupes [MWCE07] ;
- la modélisation des émotions avec le modèle Pairwise-Link-LDA axé sur le problème de la modélisation conjointe du texte et des citations [BXZ+09] ;
- l'évaluation automatique des essais : le problème de la notation automatique des essais fait l'objet de recherche depuis les années 1960 et est étroitement lié à la catégorisation automatique des textes dont *LDA* a démontré son efficacité dans la résolution des tâches de recherche, de filtrage et de classification d'information [KMS06] ;
- l'anti-Hameçonnage : les e-mails d'hameçonnage ont pour objectif d'obtenir des informations sensibles telles que celles portant sur un compte bancaire, une carte de crédit et des numéros de sécurité sociale. Étant donné que les modèles de Thème latent sont des thèmes de mots qui apparaissent ensemble dans le courrier électronique, l'utilisateur peut s'attendre à ce que, dans un courrier de phishing, les mots “cliquer” et “compte” apparaissant souvent ensemble [BCP+08], relèveraient dans ce contexte du même thème.

2.3 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 auto-*

matique 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 [Bak18].

Les méthodes pour générer cette cartographie incluent les réseaux de neurones [MSC+13], la réduction de la dimensionnalité sur la matrice de co-occurrence des mots [LG14b], les modèles probabilistes [GCPT07] et la représentation explicite en fonction du contexte dans laquelle apparaissent les mots [LG14a].

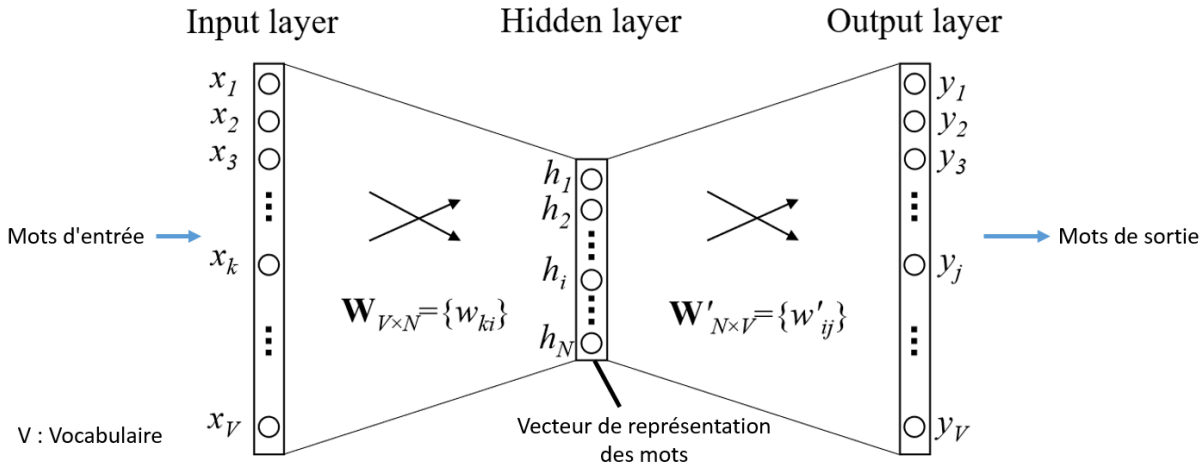


FIGURE *2.4* – ***Représentation graphique du plongement de mots***

Le modèle simple est celui du sac de mots continu (*CBOW : continuous bag-of-words*) présenté par Mikolov [MSC+13]. Dans ce modèle, similaire au modèle bi-gramme, on considère un seul mot du contexte. A partir de ce contexte, le modèle prédit un mot cible. Dans notre contexte, x_k et y_j représentent respectivement les mots d'entrée et de sortie, où $\{x_1, \ldots, x_V\}$ et $\{y_1, \ldots, y_V\}$ sont les vecteurs d'entrée et sortie du modèle, et sont de taille V représentant l'ensemble du vocabulaire. Le vecteur d'entrée est un vecteur à encodage one-hot, qui consiste à encoder une variable à V états sur V bits dont un seul prend la valeur 1 et les autres prennent la valeur 0, le numéro du bit valant 1 étant le mot d'entrée extrait du contexte. Le vecteur de sortie représente la probabilité y_j d'être le mot le plus proche pour chaque mot du vocabulaire V. Le vecteur $h = \{h_1, \ldots, h_N\}$ est appelé couche cachée et contient N neurones définis arbitrairement (cf. Figure 2.4).

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 (cf. Figure 2.5).

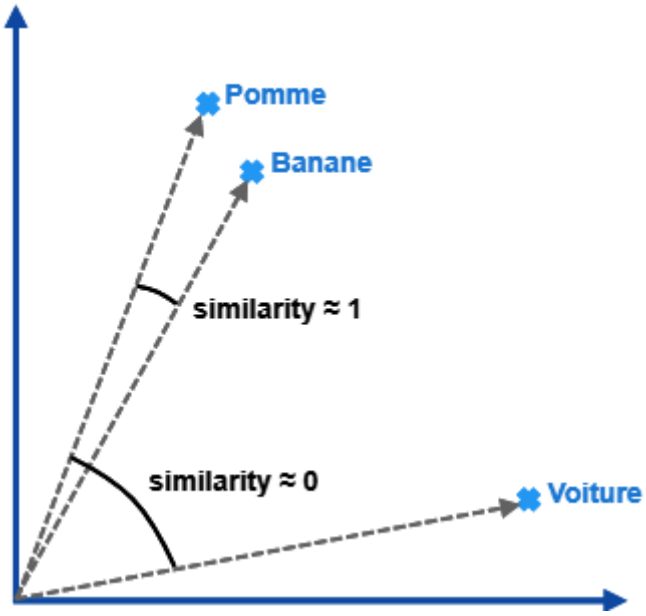


FIGURE *2.5* – ***Projection en 2 dimensions de la similarité contextuelles de mots. Une valeur de similarité proche de*** 1 ***indique des mots utilisés dans des contextes similaires, inversement,*** 0 ***pour des mots avec un faible lien contextuel.***

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 [Köh16, BGL14, SPW+13], la détection de métaphore [TBG+14, TFL+15], la reconnaissance d'entités nommées [TRB10, CWB+11], l'analyse des sentiments [SLMJ15, SBMN13] et la détection de paraphrase [BCE16, BG18].

La figure 2.6 illustre la capacité du modèle à organiser les mots dans l'espace en fonction de leur signification sémantique.

2.4 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 d'ajout de nouveaux documents au modèle lors du processus d'estimation [KMST08].

Ceci nous paraît en effet essentiel pour permettre la détermination d'un thème représentatif du *signal faible* de part ses propriétés de nouveauté et d'anormalité. On peut citer comme

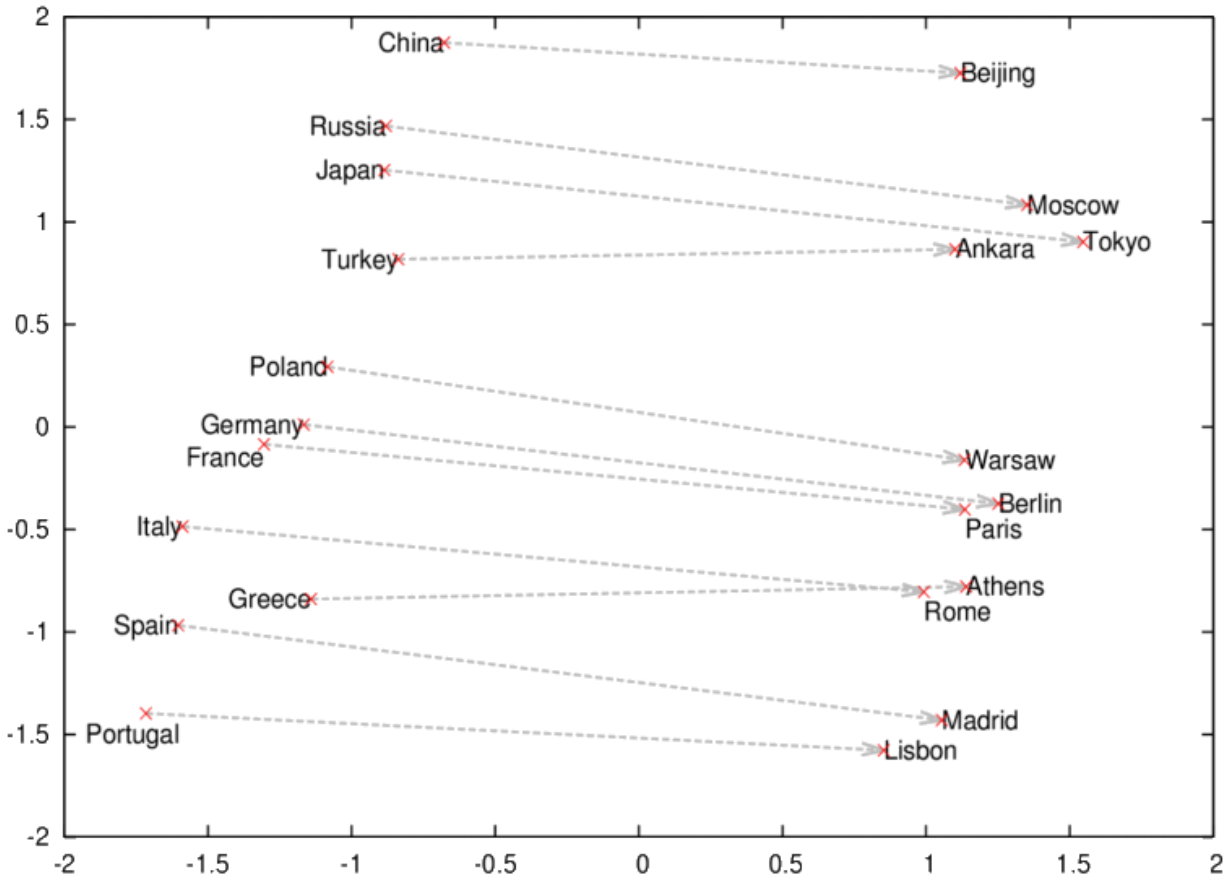


FIGURE *2.6* – ***Exemple de termes (pays et leur capitale) projetés en deux dimensions par*** Analyse en composantes principales ***(***PCA***) appliquée à leurs descripteurs*** Word2Vec***. Pendant l'apprentissage, aucune information n'est fournie sur le concept de capital. Le modèle est entrainé sur un corpus de Google News. Inspiré du travail de Mikolov [MSC+13, MCCD13]***

exemple le cas de la presse [1]qui 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* [3]. Puisque l'objectif est la détection d'un thème relatif au *signal faible*, celui-ci est par définition disjoint des autres (unitaire et non relié sémantiquement aux autres thèmes i.e. à des paradigmes existants), et selon la définition de Ah-Pine [APLB05] relativement orthogonale aux autres informations

1. Par exemple, 70 000 documents confidentiels sur les opérations de la coalition internationale en Afghanistan ont été diffusés par le site WikiLeaks [2]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.
2. `https://wikileaks.org/`
3. Ce dernier est un modèle générique de hiérarchie de thèmes 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 [BGJT04]. Ainsi les thèmes sont tous liés à travers l'arbre.

contenus dans le corpus.

2.5 Justification d'une approche conjointe modélisation thématique/plongement lexical

Toutes les approches de modélisation thématique, qu'elles soient de nature heuristique, possibiliste, probabiliste, ou floue (construites par exemple à partir de fonctions objectives) souffrent de la même difficulté : le nombre de thèmes 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 thème qui peut être de forme et de taille quelconque. De plus, les données du problème sont entachées de bruits (outliers) et d'ambiguïté entre thèmes [GMP10, Mén01, ME02]. 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 thèmes 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 thèmes recherchés très important, suivi d'une évaluation heuristique, afin d'agréger les thèmes. Un nombre excessif de thèmes 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, [ZCP+15] 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 modélisation thématique de type possibiliste ou floue 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 thèmes 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 thème, nous proposons donc de nous 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 lexi-

cal. 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 thèmes, 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 thèmes. Cette arborescence est finalement simplifiée et élaguée grâce au critère de cohérence, et seuls les thèmes cohérents au sens de notre définition des *signaux faibles* sont alors retenus. Nous appelons cette solution une modélisation thématique multi-niveaux.

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 2.7.1 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 2.7.2), nous utiliserons un réseau disponible en ligne et appris sur l'encyclopédie Wikipédia. Dans le cas de l'exemple applicatif cité en introduction sur la diffusion de documents par un lanceur d'alerte, 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.

2.6 *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 thèmes les plus cohérents parmi ceux fournis

par *LDA* et gérés à différents niveaux K. Afin de faciliter la compréhension de l'approche, nous illustrons les différentes étapes qui seront abordées plus en détail par la suite :

— La figure 2.7 illustre schématiquement la manière dont 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 précisée en détail dans la section 2.7 au travers des différents tests sur des données artificielles et sur des données proches du réel ;

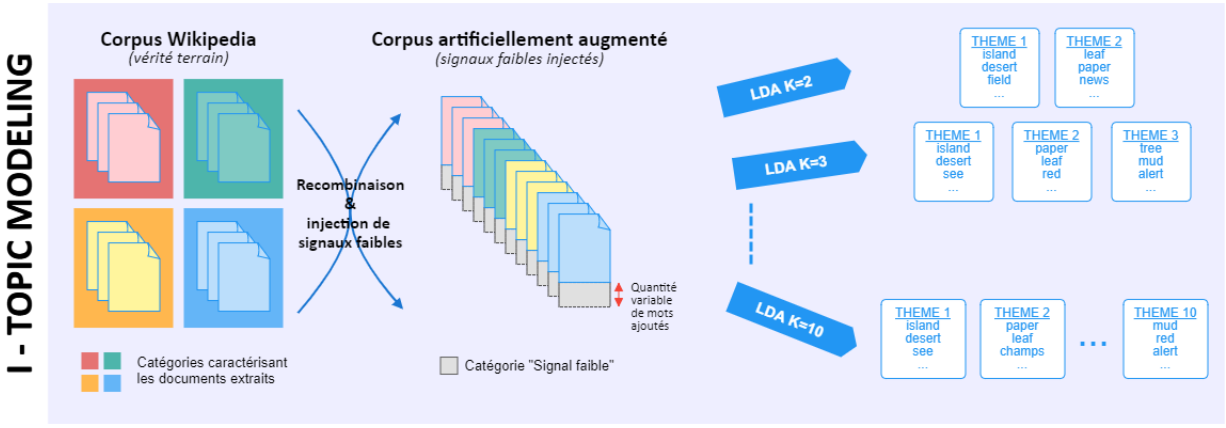


FIGURE *2.7* – ***Etape de génération du corpus de test sur la base d'un jeu de données extrait de Wikipédia. Nous appliquons*** LDA ***sur l'ensemble des documents, tout en faisant varier le nombre de thèmes.***

— Pour l'analyse des documents et l'extraction du *signal faible*, nous adoptons une approche conjointe *LDA*/*Word2Vec* (illustrée sur les figures 2.7 et 2.8) appelée modélisation thématique multi-niveaux. Nous utilisons *Word2Vec* pour définir un critère de cohérence sur les thèmes obtenues afin de qualifier la similitude sémantique des ensembles de mots présents dans chacun d'entre eux. Le modèle *Word2Vec* est pré-entrainé sur le corpus français de Wikipedia.

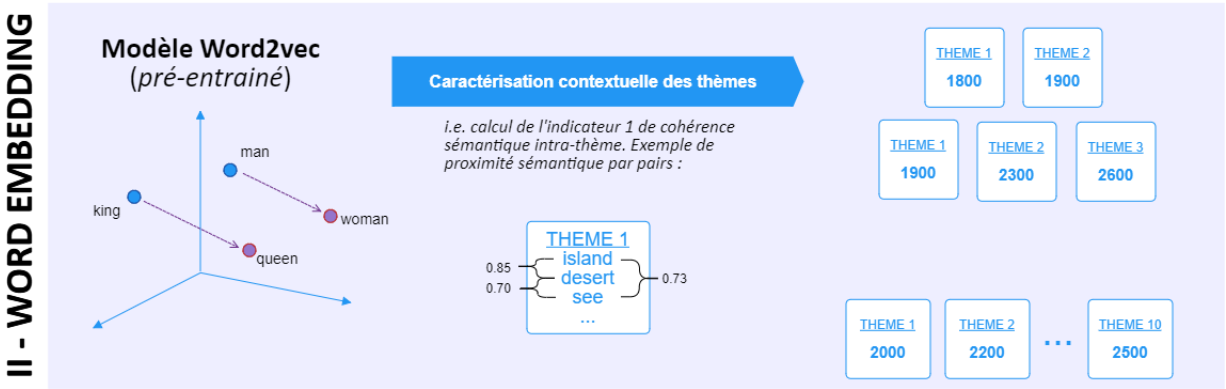


FIGURE *2.8* – ***Etape de caractérisation contextuelle des thèmes sur la base d'un modèle*** Word2Vec ***pré-entrainé.***

— L'ensemble des thèmes est relié entre-eux sous la forme d'une arborescence via un critère

de ressemblance. Celle-ci est élaguée (cf. Figure 2.9) afin de dégager un sous-ensemble de thèmes où au moins l'un d'entre-eux est susceptible de contenir les mots-clés du *signal faible*.

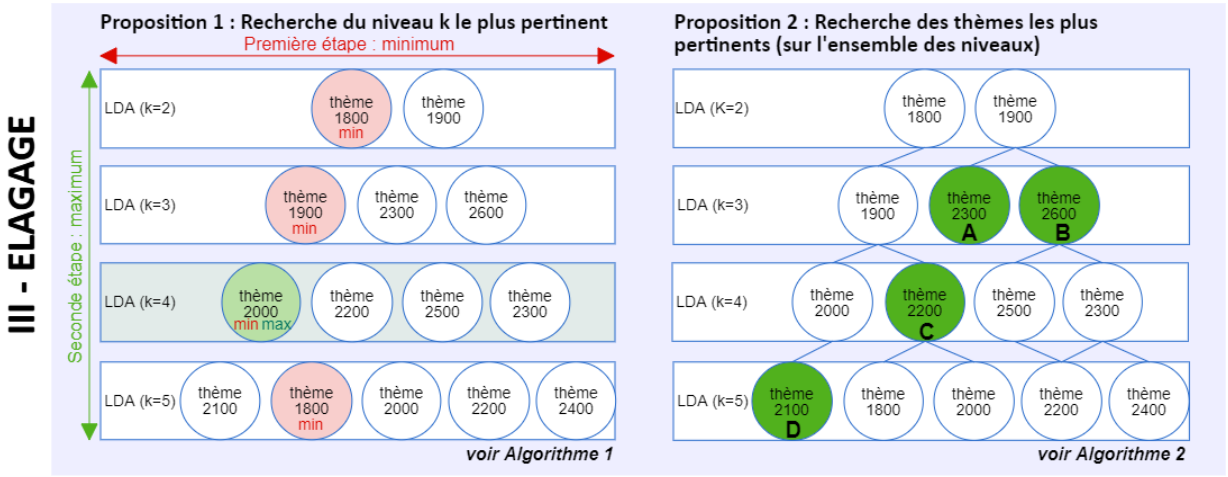


FIGURE *2.9* – ***La proposition 1 montre la recherche du niveau de l'arborescence donnant les thèmes les plus cohérents au sens du critère de ressemblance. La proposition 2 montre l'élagage des thèmes reliés entre eux dans l'arborescence au moyen d'un critère de ressemblance. L'objectif est dans ce deuxième cas de rechercher les thèmes les plus cohérents sur l'ensemble de l'arborescence.***

— Enfin sur la figure 2.10, les thèmes porteurs de *signaux faibles* potentiels sont identifiés. La méthode de pondération *tf-idf* est utilisée pour obtenir les mots-clés pertinents au sens de la définition d'un *signal faible*.

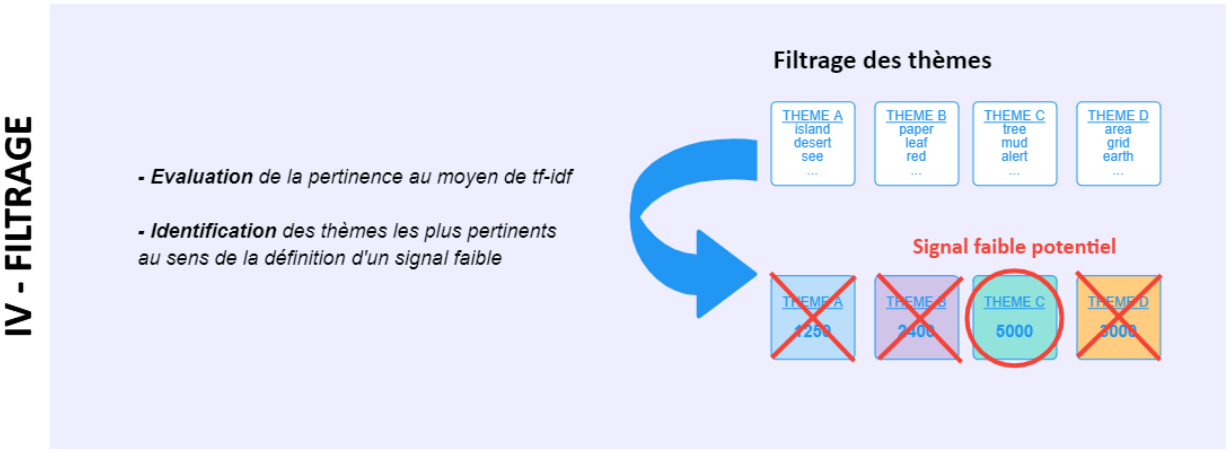


FIGURE *2.10* – ***Filtrage des thèmes au moyen d'une évalutation de la pertinence au moyen de*** tf-idf. ***Nous identifions les thèmes*** signaux faibles ***potentiels.***

2.6.1 Cas d'utilisation de *LDA* sur Wikipédia

Pour illustrer l'utilisation de *LDA*, nous nous concentrons sur un sous-ensemble de documents extrait de la version française de Wikipedia (snapshot du 08/11/2016) sur 5 Thèmes

différents : Economie, Histoire, Informatique, Médecine et Droit. La figure 2.11 donne les caractéristiques des Thèmes relatifs aux documents quant au nombre de mots rencontrés respectivement dans un seul Thème, deux Thèmes ou au moins trois Thèmes. 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. On trouve des feuilles communes sur plusieurs branches. Bairi [BCR15] explique que, dans Wikipédia, une catégorie principale couvre tous les documents présents dans l'encyclopédie si nous explorons ses sous-catégories à une profondeur de 10. En outre, il y a de forts chevauchements entre les catégories, ce qui explique pourquoi il y a certainement peu de pages spécifiques à un Thème particulier. La répartition des mots des documents est donnée sur la figure 2.11.

	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 *2.11* – ***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 2.12 illustre l'utilisation de la méthode *LDA* pour laquelle nous devons spécifier le nombre de thèmes K. En sortie, nous obtenons une classification en Thèmes triés par ordre de pertinence.

FIGURE *2.12* – ***Mise en œuvre de*** LDA ***sur un corpus de documents. En sortie, nous obtenons un ensemble de thèmes dont le nombre est spécifié en paramètre d'entrée.***

LDA est une méthode de modélisation thématique non supervisée et n'associe pas un label aux thèmes trouvés. Dans le cas d'un grand nombre de documents, il est difficile d'estimer *a priori* le nombre de thèmes 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 2.1 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 thème 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 ville roi nom église ...	film album premier années ans ...	saison club première premier tour ...	guerre pays france général français ...

TABLE *2.1* – ***Liste des 5 premiers mots de chaque thème***

2.6.2 Indicateur de cohérence en tant que mesure intra-thème

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 (thèmes) dans le contexte du corpus de documents. Ce premier indicateur est défini comme suit :

$$I_1 = \sum_{w \epsilon E} w2vSim(w_i, w_j) \tag{2.1}$$

où I_1 représente la somme des valeurs de similarité de toutes les combinaisons de paires de mots dans chaque thème, avec $E = \{w_1, w_2, \ldots, w_L\}$ l'ensemble des L premiers mots supportant le thème. $w2vSim$ représente la mesure de similarité définie dans *Word2Vec* comme la similarité cosinus entre deux vecteurs [MCCD13]. Plus la valeur de I_1 est grande, plus le thème contient de mots régulièrement employés ensemble. Dans les expérimentations, nous choisissons $L = 100$ qui 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 $L = 10$ mots pour les tests effectués sur les corpus "proches" du réel (cf. section 2.7.2).

Nous proposons d'utiliser cet indicateur sur les thèmes 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 thèmes découverts lors des différentes exécutions dans un ordre aléatoire. Une étape supplémentaire présentée dans la section 2.6.4 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 2.6.3) visant à rechercher le nombre de thèmes (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 2.6.4) qui, de manière plus avancée, par une analyse approfondie de l'arborescence, combine les meilleurs thèmes renvoyés par *LDA* sur toutes les partitions (ou valeurs de k testées).

2.6.3 Recherche du paramètre k conduisant aux thèmes les plus pertinents au sens du critère de cohérence

L'algorithme 1 consiste à rechercher le niveau de l'arborescence donnant les thèmes 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 thèmes présents sur ce niveau. Le niveau k choisi (et donc le nombre de thèmes pertinents au sens du critère) correspond à celui dont la valeur minimale est la plus élevée (cf. Eq. 2.2 et Figure 4).

$$\underset{k}{Argmax}(min_{c_l}(I_1(c_l))), c_l \in \{c_1 \ldots c_k\}, k \in \{2 \ldots K\} \tag{2.2}$$

Concernant l'algorithme, le résultat obtenu par la variable *meilleurk*, désigne au sens du critère contextuel, I_1, le niveau k où se situent les thèmes les plus pertinents.

Data : P = Liste des nombres de thèmes 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 $c_l \in \{c_1 \dots c_k\}$ **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

Algorithme 1 : ***Recherche du niveau de l'arborescence donnant les thèmes les plus cohérents***

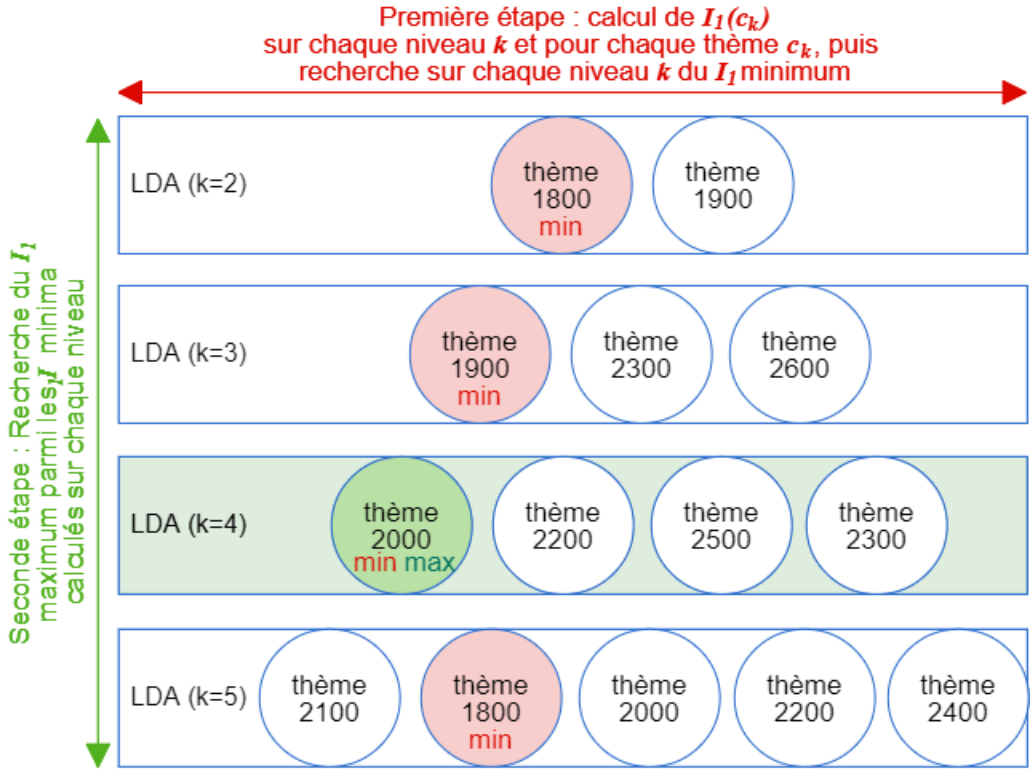


FIGURE *2.13* – ***Représentation de l'application de l'algorithme 1 sur les différentes partitions*** LDA ***obtenues pour*** $k \in \{2 \dots 5\}$. ***Les partitions sont organisées par niveau : la partition obtenue pour*** $k = 2$ ***est au niveau le plus haut. Les niveaux sont représentés par des rectangles. Les valeurs pour chaque thème sont celles de l'indicateur*** I_1. ***Les cercles rouge et vert représentent les plus petites valeurs de chaque niveau de*** LDA. ***Parmi ces valeurs, la plus élevée correspond à*** LDA 4 ***(*** $k = 4$***), désignant la partition la plus cohérente au sens du critère contextuel.***

2.6.4 Une approche heuristique pour déterminer les thèmes 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 thèmes de niveaux adjacents. Celui-ci est calculé en utilisant un indicateur de ressemblance

utilisant la distance de Bhattacharyya définie comme suit [Bha43] :

$$I_2 = \sum_{w \epsilon E} \sqrt{p_{w_i} \cdot q_{w_i}} \tag{2.3}$$

Pour l'ensemble E défini par les mots w présents dans un thème c_l de niveau k et un thème 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 thèmes respectifs c_l et c_{l+1}.

Il est alors possible d'extraire sur toute l'arborescence les thèmes 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 thèmes de niveau respectif k et $k+1$ (voir tableau 2.2). 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 thème nouvellement rencontré, retenu comme pertinent, conduit au retrait dans l'arborescence de tous les thèmes parents et fils. Les relations entre les thèmes (décrites par l'indicateur I_2) ne sont prises en compte qu'au-delà d'un seuil arbitrairement défini (cf. Figure 2.14). L'algorithme 2 formalise cette heuristique où les fonctions $Parents(c_l)$ et $Fils(c_l)$ récupèrent respectivement la liste des thèmes parent et fils du thème c_l. Nous obtenons alors une liste des thèmes pertinents non 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**
  meilleurTheme $\leftarrow$ `Max(`T`)`;
  themesRetenus $\longleftarrow$ themesRetenus + {meilleurTheme };
  for $t \in$ `Parents(meilleurTheme)` **do**
    $T \longleftarrow T - t$;
  end
  for $t \in$ `Fils(meilleurTheme)` **do**
    $T \longleftarrow T - t$;
  end
end
return themesRetenus

Algorithme 2 : ***Récupération des thèmes pertinents dans l'arborescence*** LDA

L'algorithme 2 conduit aux propriétés suivantes :

$$\begin{cases} \overline{I_1(c_i)}^{Alg2} \geqslant \overline{I_1(c_i)}^{LDA_k} & \forall\ k \\ \underset{i}{max} I_1^{Alg2}(c_i) \geqslant \underset{i}{max}\ I_1^{LDA_k}(c_i) & \forall\ k \end{cases} \tag{2.4}$$

La moyenne des cohérences sémantiques des thèmes obtenus est augmentée.

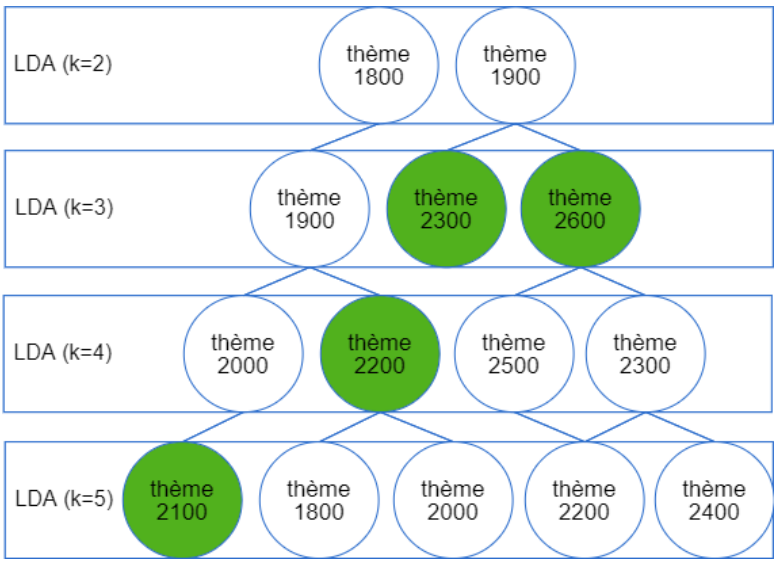


FIGURE *2.14* – ***Représentation de l'algorithme 2 sur les différentes partitions*** LDA ***obtenues pour*** $k \in \{2 \dots 5\}$ ***après avoir construit l'arborescence en utilisant l'indicateur de ressemblance*** I_2***. Un lien est formé entre chaque thème de niveau*** k ***et un thème de niveau*** $k+1$ ***si la valeur de l'indicateur*** I_2 ***est supérieure à un seuil défini arbitrairement. Les partitions sont organisées par niveau : la partition obtenue pour*** $k = 2$ ***est au niveau le plus haut. Les niveaux sont représentés par des rectangles. Les thèmes ont tous été triés dans une liste par ordre croissant selon l'indicateur I1 sans distinction du niveau*** k ***(cf. Tableau 2.2). Nous sélectionnons le premier élément de la liste et supprimons ses fils et parents avec les méthodes*** $Parents(c_k)$ ***et*** $Fils(c_k)$***. Cette tâche est exécutée jusqu'à ce que la liste soit vide. Les éléments récupérés forment les thèmes les plus pertinents selon l'algorithme 2.***

2.7 Expérimentations

Une preuve de concept est réalisée pour évaluer l'approche présentée. A cette fin, nous créons 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. Dans ces travaux, nous utilisons le mot "thème" pour définir les résultats de *LDA* et le mot "Thème" pour définir les ensembles de mots utilisés comme vérité terrain. Ces ensembles de données permettent une analyse objective et sans ambiguïté. Chaque corpus comporte plusieurs Thèmes. Ils contiennent des documents considérés comme des mélanges de ces Thèmes. Dans la base de données "synthétique" 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*.

LDA	thème	Valeur
$LDA(k=3)$	3	2600
$LDA(k=4)$	3	2500
$LDA(k=5)$	5	2400
$LDA(k=3)$	1	2300
$LDA(k=4)$	4	2300
$LDA(k=4)$	2	2200
$LDA(k=5)$	4	2200
$LDA(k=5)$	1	2100
$LDA(k=4)$	1	2000
$LDA(k=5)$	3	2000
$LDA(k=2)$	2	1900
$LDA(k=3)$	1	1900
$LDA(k=2)$	1	1800
$LDA(k=5)$	2	1800

TABLE *2.2* – ***Liste des thèmes triés par valeur de l'indicateur*** I_1 ***selon l'exemple en figure 2.14. La sélection des thèmes résultats est faite par suppression des éléments de la liste selon les relations entre les thèmes décrits par l'indicateur*** I_2***. Tant que la liste est non vide, le premier élément est sélectionné comme thème résultat et ses fils et parents (avec les méthodes*** $Parents(c_k)$ ***et*** $Fils(c_k)$***) sont supprimés de la liste selon l'arborescence***

2.7.1 Tests 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
- Thème 5 "mots-outils" : nombres de 600 à 699
- Thème 6 *signal faible* : nombres de 900 à 999

2.7.1.1 Base de documents

Notre base de documents initiale pour la réalisation des tests comprend pour chacun des 4 Thèmes principaux 50 documents, ainsi que 50 autres documents associés au Thème *signal faible*. Le nombre de documents est fixé arbitrairement. La base de document servira également pour l'inférence à partir des thèmes. Chacun des documents générés puise ses 10 000 mots dans

le jeu de données décrit par le tableau 2.3. Un exemple de document est présenté sur la figure 2.15.

	Nombre de documents	**Mots de 0 à 99**	**Mots de 100 à 199**	**Mots de 200 à 299**	**Mots de 300 à 399**	**Mots de 900 à 999**
Document du Thème 1	50	10 000	0	0	0	0
Document du Thème 2	50	0	10 000	0	0	0
Document du Thème 3	50	0	0	10 000	0	0
Document du Thème 4	50	0	0	0	10 000	0
Document du Thème 6 ***signal faible***	50	0	0	0	0	10 000

TABLE *2.3* – ***Description du jeu de données. Dans le corpus initial, chaque document est composé intégralement de mots appartenant au champ lexical d'un même Thème.***

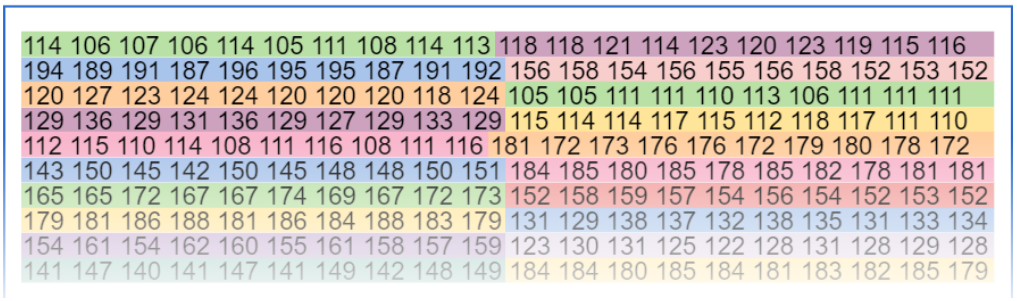


FIGURE *2.15* – ***Échantillon d'un document. Les couleurs représentent les groupes de mots. Le document contient des mots du Thème principal*** 2

Ainsi, dans ce corpus initial, chaque document est composé intégralement de mots appartenant au champ lexical d'un même Thème. Nous présentons par la suite plusieurs expérimentations utilisant des jeux de données différents construits à partir de celui-ci. Dans les différents tests effectués, plusieurs valeurs de seuil ont été expérimentées (entre $0,4$ et $0,9$) pour l'indicateur I_2 (Eq. (2.3)). Cette valeur possède un impact faible sur la qualité et variabilité des résultats. Nous choisissons la valeur $0,7$ dans la suite de nos expérimentations.

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

— pour chaque thème 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 considéré comme détecté lorsque la valeur du Thème 6 est la plus importante [4].

4. Par la suite, il sera nécessaire de déterminer parmi les clusters obtenus celui ou ceux qui portent le *signal*

Pour chaque résultat présenté, et pour chaque algorithme, nous effectuons 10 lancés sur des corpus propres à chaque exécution. Le but étant d'évaluer la robustesse des algorithmes vis-à-vis de la variabilité des corpus et du non-déterminisme de *LDA*. On mesure le nombre de détection du Thème *signal faible* parmi l'ensemble des thèmes en fonction du nombre total de lancement (10 dans notre cas). A titre comparatif, les différents tests sont également effectués en utilisant plusieurs *LDA* indépendants (avec k variant de 5 à 10).

La figure 2.16 illustre comment les 4 corpus de tests sont construits.

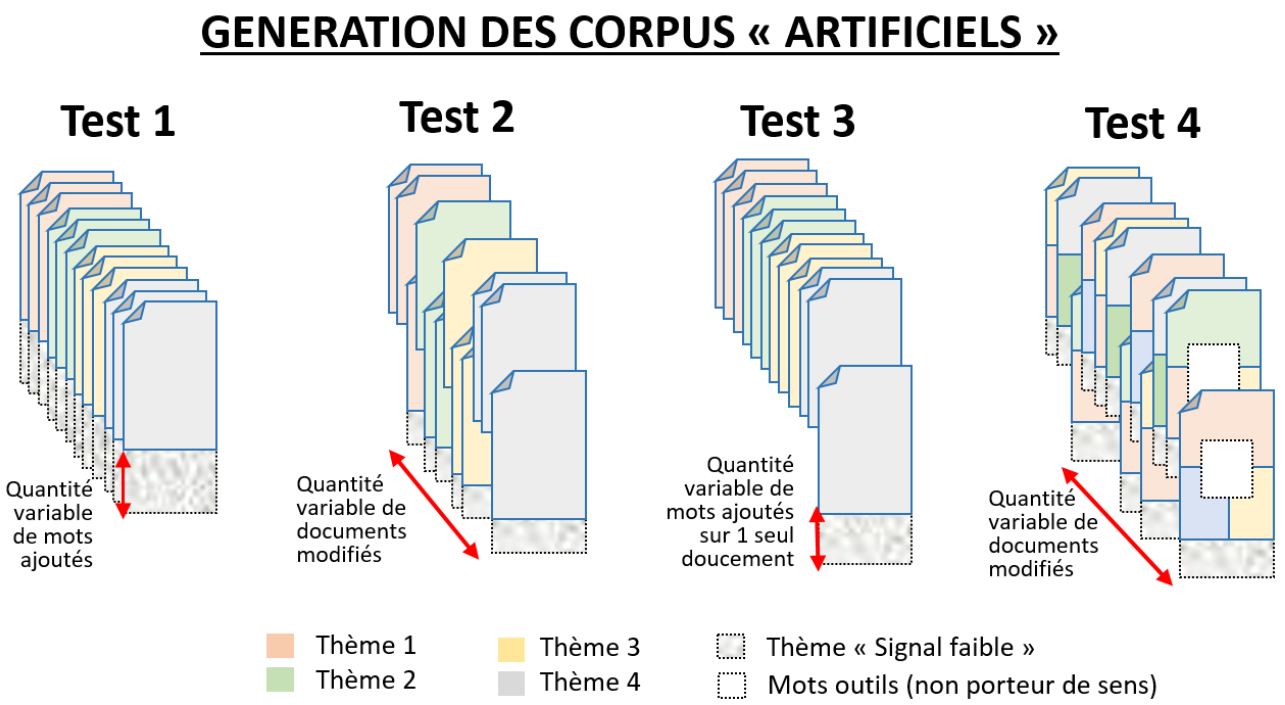


FIGURE *2.16* – ***Construction des 4 corpus de tests artificiels. Les mots du Thème signal faible sont injectés en quantité variable dans chaque document du corpus (test 1), ou bien dans une quantité variable de documents (test 2), ou encore en quantité variable dans un seul document avec (test 3) ou sans mots-outils (test 4)***

.

2.7.1.2 Test 1 : chaque document est composé de 10 000 mots relevant d'un Thème principal et d'un nombre variable de mots du Thème *signal faible*

Dans ce test, nous faisons évoluer le nombre de mots du Thème *signal faible* intégrés dans chaque document. Ces derniers sont donc composés de 10 000 mots relevant d'un Thème principal et d'un nombre variable de mots du Thème *signal faible*. L'objectif est d'évaluer l'impact de ce nombre sur les résultats de l'algorithme 2 par rapport à *LDA* seul.
Dans le cadre de ce test, l'évolution du nombre de mots du Thème *signal faible* s'effectue par pas de 100 mots, de 0 à 2500.

faible

SecretWords	Algorithme 2		LDA			
			5	6	7	8
0	0		0	0	0	0
100	1		0	2	3	5
200	6		0	1	1	0
300	3		0	1	1	2
400	6		0	2	4	2
500	4		0	0	2	3
600	9		2	1	1	4
700	7		1	5	3	4
800	9		2	4	2	2
900	7		1	0	1	5
1000	7		2	4	6	1
1100	7		2	2	5	5
1200	9		3	3	2	2
1300	8		1	1	3	3
1400	9		0	2	3	6
1500	6		0	6	4	4
1600	6		2	2	4	5
1700	8		0	3	6	4
1800	8		2	6	3	7
1900	6		5	5	3	5
2000	9		2	2	2	3
2100	9		1	3	2	5
2200	9		2	1	4	4
2300	9		2	3	6	4
2400	7		5	3	1	6
2500	9		2	2	2	4
Total	178		37	64	74	95
Réussite	71%		15%	26%	30%	38%

FIGURE *2.17* – ***Expérimentation 1 : résultats de l'application de l'algorithme 2 comparativement aux 6*** LDA ***seuls paramétrés avec des valeurs de*** K ***allant de 5 à 8. Sur chaque ligne est donné le nombre de mots,*** $\#w$***, relatif au*** signal faible ***inséré dans chaque document "SecretWords-*** $\#w$***".***

Le nombre figurant dans chaque case correspond au nombre de fois où le thème relatif au *signal faible* est détecté (10 fois au maximum). Les résultats présentés dans la Figure 2.17 montrent une amélioration sensible de la détection par rapport à *LDA* seul, quelque soit la valeur du paramètre K (ou niveau) de *LDA*. Les thèmes détectés relatifs au *signal faible* peuvent figurer à un niveau quelconque de l'arborescence comme le montre finalement les résultats des *LDA*. Les résultats montrent également qu'une détection du *signal faible* est fort probable à partir d'une proportion de mots du Thème *signal faible* dans chaque document de 600/10600, soit environ 5%.

2.7.1.3 Test 2 : chaque document est composé de 10000 mots relevant du Thème principal ; une partie variable d'entre-eux contient en plus des mots relatifs au *signal faible*

L'expérience présentée dans cette section a pour objectif d'étudier l'influence du nombre de documents porteurs du *signal faible*.
Dans cette expérimentation, chaque document généré contient 10 000 mots provenant d'un Thème principal. Il y a bien sûr toujours 4 Thèmes représentés. Une part variable de documents contiendra en plus 2 500 mots relatifs au Thème *signal faible*. Le nombre de documents porteurs du *signal faible* varie de 0 à 200 par pas de 10. Les documents non porteurs du *signal faible* correspondent au complément pour obtenir un jeu de 200 documents au total construits sur les 4 Thèmes principaux. Des tests complémentaires ont été effectués pour un nombre de documents porteurs entre 0 et 10 et 190 et 200. Le nombre de mots appartenant au Thème relatif au *signal faible* est ici de 2 500.

On constate sur le tableau de la Figure 2.18 que l'algorithme 2 détecte correctement le thème porteur du *signal faible* avec un taux de 99.7%. Il apporte une amélioration entre 6% et 24.7% selon le paramètre k utilisé pour les *LDA* seuls. Ici aussi, le tableau montre que le thème porteur du *signal faible* peut se situer à un niveau quelconque de l'arborescence.

2.7.1.4 Test 2 Bis : similaire à l'expérimentation précédente mais avec une proportion de 5% de mots du Thème *signal faible* par document porteur

Lors de l'expérimentation 1 présentée sur la figure 2.17, nous avions remarqué une détection significative à partir de 600 mots du Thème *signal faible* injectés par document. Nous appliquons le même protocole que celui du test 2, mais choisissons de n'insérer que 600 mots du Thème *signal faible* au lieu des 2500 initiaux, soit une réduction de 75%, et donc une proportion de 5% seulement par document.

Les résultats présentés sur la Figure 2.19 montrent à nouveau la robustesse de l'approche proposée, même en cas de diminution significative de la quantité de mots relatifs au *signal*

DocWithSecretWords	Algorithme 2		LDA			
			5	6	7	8
0	0		0	0	0	0
1	10		2	2	4	6
2	10		6	10	9	10
3	10		9	9	10	10
4	10		6	10	10	10
5	10		10	10	10	10
6	10		8	9	10	10
7	10		9	9	10	10
8	10		10	9	10	10
9	10		10	10	10	10
10	10		9	10	10	10
20	10		10	10	10	10
30	10		10	10	10	10
40	10		10	10	10	10
50	10		9	10	10	10
60	10		8	10	10	9
70	10		10	10	10	10
80	10		10	10	10	10
90	10		8	10	10	10
100	10		8	10	10	10
110	10		10	10	10	10
120	10		9	9	10	9
130	10		6	10	10	10
140	10		9	10	10	10
150	10		8	9	10	10
160	10		9	10	10	10
170	10		10	10	10	10
180	10		8	10	10	10
190	10		9	9	10	10
191	10		5	7	8	9
192	10		6	6	9	10
193	10		4	8	8	7
194	10		6	6	8	6
195	10		7	8	8	8
196	10		4	5	7	7
197	10		2	3	5	6
198	10		3	6	1	6
199	9		3	4	6	7
200	10		5	1	7	6
Total	379		285	319	340	346
Réussite	99.7%		75.0%	83.9%	89.5%	91.1%

FIGURE *2.18* – ***Expérimentation 2 : résultats de l'application de l'algorithme 2 comparativement aux 6*** LDA ***seuls paramétrés avec des valeurs de*** K ***allant de 5 à 8. Sur chaque ligne est décrit le nombre de documents,*** $\#d$***, porteurs du*** signal faible ***(i.e. contenant 2500 mots relatifs au Thème*** signal faible***).***

faible injectés dans les "documents porteurs". Le score de l'algorithme 2 dans le test 1 figurant à la ligne "SecretWords-600" correspond à celui de la ligne "DocWithSecretWords-200" de la

DocWithSecretWords	Algorithme 2		LDA			
			5	6	7	8
0	0		0	0	0	0
10	10		9	10	10	10
20	10		10	9	10	8
30	10		8	10	10	10
40	10		4	8	10	9
50	10		5	9	10	10
60	10		7	8	9	10
70	10		3	8	10	9
80	10		7	9	9	9
90	10		4	8	9	10
100	10		7	8	9	9
110	10		4	7	10	10
120	10		4	6	10	9
130	10		6	9	10	9
140	10		5	8	5	7
150	10		3	6	7	6
160	10		4	7	5	8
170	10		4	6	9	9
180	10		5	4	6	9
190	10		3	3	6	7
200	10		0	2	2	0
Total	200		102	145	166	168
Réussite	100%		51%	73%	83%	84%

FIGURE *2.19* – ***Expérimentation 2 Bis : résultats de l'application de l'algorithme 2 comparativement aux 6*** LDA ***seuls paramétrés avec des valeurs de*** K ***allant de 5 à 8. Sur chaque ligne est décrit le nombre de documents,*** $\#d$***, porteurs du*** signal faible ***(i.e. contenant 600 mots relatifs au*** signal faible***).***

figure 2.19. La différence de détection provient du non déterministe de l'algorithme *LDA*. Une amélioration de l'ordre de 16% à 49% est ici constatée selon le paramètre k utilisé par *LDA* seul.

2.7.1.5 Test 3 : 1 seul document est porteur du *signal faible* parmi les 200 documents du corpus. Il contient un nombre variable de mots appartenant au Thème *signal faible*

L'expérience présentée dans cette section vise à étudier l'influence de la quantité de mots relevant du Thème *signal faible* injectés dans le document porteur. De manière similaire au premier test (Cf. Section 2.7.1.2), le nombre de mots varie de 0 à 2 500 mots par pas de 100 mots.

SecretWords	Algorithme 2	LDA						
		2	3	4	5	6	7	8
0	0	0	0	0	0	0	0	0
100	5	0	0	0	0	0	1	4
200	8	0	0	0	1	1	5	6
300	8	0	0	0	2	1	6	6
400	9	0	0	0	1	4	3	6
500	8	0	0	0	2	4	5	7
600	10	0	0	1	4	8	4	10
700	10	0	0	1	3	4	8	8
800	10	0	0	1	0	7	10	7
900	10	0	0	0	4	7	8	10
1000	10	0	0	0	4	6	6	9
1100	10	0	0	0	5	4	9	7
1200	10	0	0	0	4	7	6	9
1300	10	0	0	0	5	6	8	10
1400	10	0	0	0	7	5	9	8
1500	10	0	0	1	5	5	8	8
1600	10	0	0	0	4	8	8	8
1700	10	0	0	0	2	8	9	10
1800	10	0	0	1	7	8	10	10
1900	10	0	0	1	3	9	6	8
2000	10	0	0	0	3	5	8	9
2100	10	0	0	0	3	7	7	9
2200	10	0	0	0	2	7	10	9
2300	10	0	0	1	7	6	7	7
2400	10	0	0	0	6	9	9	8
2500	10	0	0	0	8	7	8	6
Total	238	0	0	7	92	143	178	199
Réussite	95%	0%	0%	3%	37%	57%	71%	80%

FIGURE *2.20* – ***Expérimentation 3 : résultat de l'application de l'algorithme 2 comparativement aux 9*** LDA ***seuls paramétrés avec des valeurs de*** K ***allant de 2 à 8. Sur chaque ligne est donné le nombre de mots,*** $\#w$***, appartenant au Thème relatif au*** signal faible ***dans le document "SecretWords-*** $\#w$***".***

Les résultats présentés dans la Figure 2.20 montrent des scores bien supérieurs pour l'algorithme 2. On note une amélioration de 7% à 58%. Le thème relatif au *signal faible* est donc détecté de manière robuste, même si un seul document est porteur et même s'il présente un taux faible, 5%, de mots appartenant à ce Thème.

2.7.1.6 Test 4 : chaque document présente un Thème principal et deux Thèmes secondaires ainsi qu'un Thème correspondant à des mots-outils. Une partie variable d'entre-eux contient en plus des mots relatifs au *signal faible*

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

TABLE *2.4* – ***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 2.4. 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.

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 de la Figure 2.21 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, secondaire 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%).

Il n'est pas possible de s'assurer qu'une valeur de k donnée de *LDA* permet de détecter correctement le thème relatif au *signal faible*. Même s'il est détecté au niveau k, sa cohérence au sens du critère donné (défini dans la section 2.6.2) 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 thèmes qui ne sont pas nécessairement représentatifs des Thèmes de tous les documents (i.e. on assiste à une sursegmentation). Même si le thème relatif au *signal faible* fait partie de l'un des thèmes 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 thèmes les plus cohérents (au sens du critère). Ces thèmes 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.

2.7.2 Tests sur des corpus de données réelles

2.7.2.1 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 Thèmes 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.

Leur récupération est réalisée par extraction de tous les articles depuis le fichier XML du dump. Ensuite, le parcours et la récupération des Thèmes liés et articles associés se font au moyen d'une base de données SQL. Enfin, nous déplaçons les articles identifiés durant la phrase de parcours de la base de données pour être utilisés comme corpus de test durant les expérimentations.

Pour réaliser le test, nous disposons d'un corpus de documents relatifs à 5 Thèmes :

— Economie : 44 876 documents
— Histoire : 92 041 documents
— Informatique : 25 408 documents
— Médecine : 22 143 documents

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%

FIGURE *2.21* – ***Expérimentation*** 4 ***: 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.***

— Droit : 9 964 documents

Pour le besoin de notre expérimentation et pour permettre une évaluation par certaines

métriques, il est nécessaire de générer un nouveau corpus de test qui implique un *signal faible* simulé. Pour ce faire, nous avons extrait un certain nombre de mesures statistiques de notre corpus initial, puis défini trois groupes de mots, comme le montre le tableau 2.11 :

1. le groupe des mots communs, ceux appartenant à 3 Thèmes ou plus (ils représentent les 12 premiers pour cent des mots du corpus triés par occurrence) ;
2. les groupes de mots relevant de deux Thèmes ;
3. ainsi que ceux appartenant à un Thème unique.

La figure 2.22 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.

Pour ce test, nous avons utilisé un modèle *Word2Vec* pré-entrainé sur le corpus français de Wikipédia (Dump du 07/11/2016) [Sch16]. L'identification du thème *signal faible* s'effectue de la façon suivante :

— Au préalable, l'appartenance de chaque mot aux 5 Thèmes a été effectuée 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 thème 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 thème, 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* [Mai19a], *http://doi.org/10.5281/zenodo.162792* [Sch16]

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 thèmes ainsi que la meilleure détection des thèmes de *signaux faibles*. Cette valeur affecte cependant le nombre de thèmes 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.

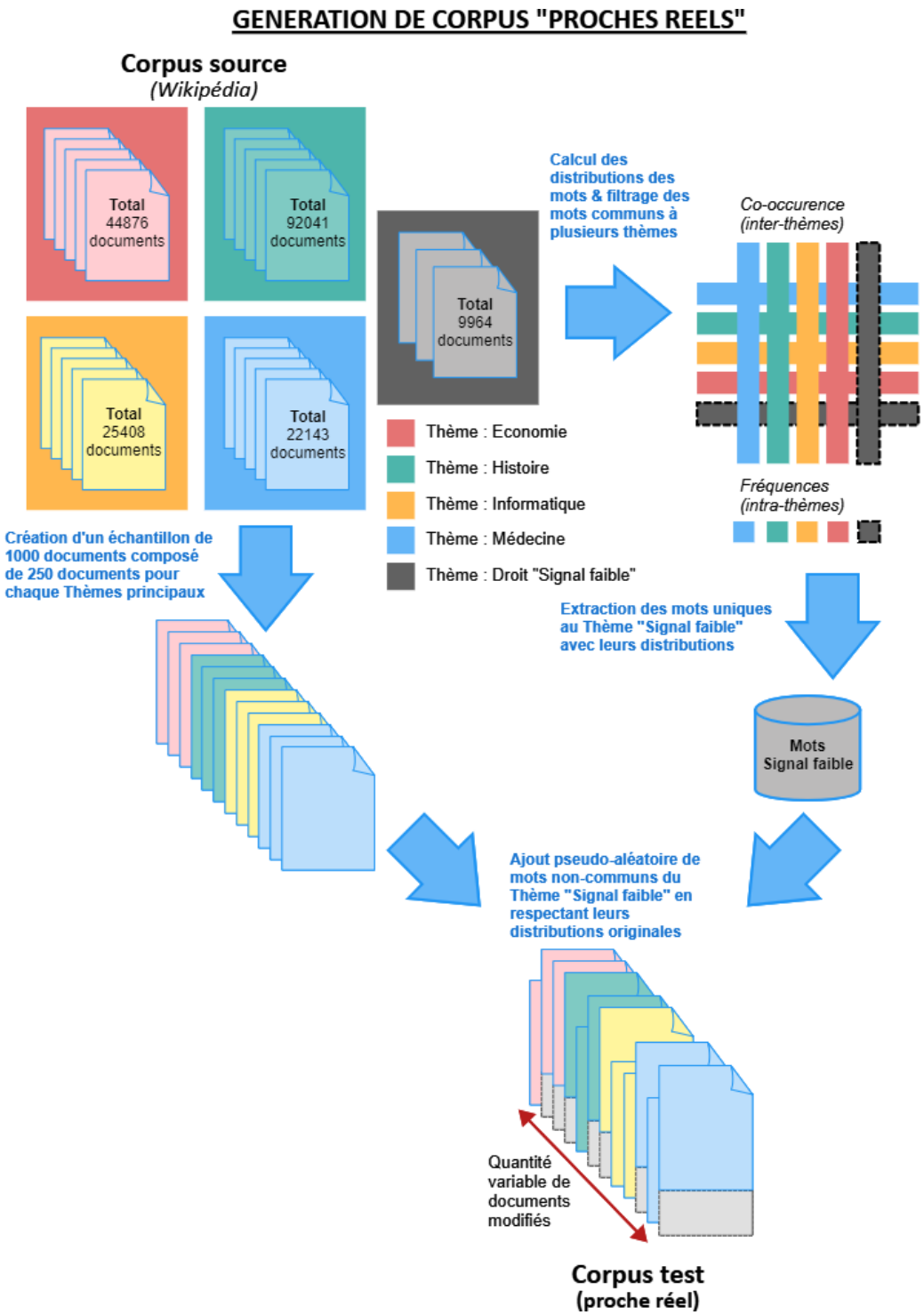


FIGURE *2.22* – ***Méthode de génération du corpus "proche du réel" qui consiste en l'injection de mots dits "non-communs" empruntés au Thème*** signal faible ***(Droit en l'occurence) en respectant leur distribution originale. Ces mots non-communs sont identifiés par une étude de co-occurrence entre Thèmes.***

Les résultats obtenus (cf. Figure 2.23), montrent la robustesse de l'algorithme 2 même pour un très petit nombre de mots injectés du Thème Droit (*signal faible*) comparée 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). Notre recherche a également porté sur la valeur de cohérence du thème *signal faible* suivant les différents niveaux de *LDA*. Dans la figure 2.24, nous montrons que l'algorithme 2 détecte le thème *signal faible* ayant le plus de cohérence dans l'arborescence (i.e. pour tous les niveaux de *LDA* $k \in \{2 \dots 8\}$). L'algorithme *LDA* utilisé seul donne parfois une partition où le thème *signal faible* est présent (avec une valeur de cohérence de similarité inférieure, cf. Figure 2.24). Ce test montre donc l'intérêt et la contribution de cette étude dans la détection d'un *signal faible* par une approche de modélisation thématique multi-niveaux.

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 thèmes trouvés.

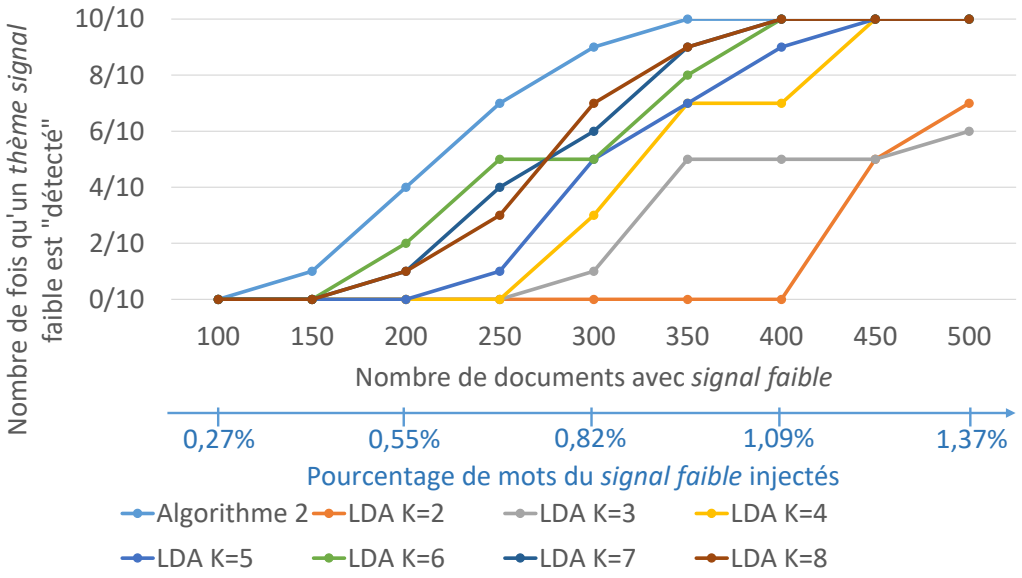


FIGURE *2.23* – ***Résultats de l'algorithme 2 comparé aux 7*** LDAs ***originaux paramétrés avec*** k ***variant de*** 2 ***à*** 8 ***sur la détection d'un thème*** signal faible ***dans les thèmes déterminés par l'approche de modélisation thématique multi-niveaux. Dans chaque document, nous insérons*** 3 ***séries de*** 4 ***mots du Thème Droit (***signal faible***). L'identification du thème*** signal faible ***est réalisée par le calcul de la somme des poids des mots n'appartenant qu'à un des 5 Thèmes (Economie, Histoire, Informatique, Médecine et Droit (***Signal faible***)). Le thème est identifié*** signal faible ***lorsque la valeur du Thème Droit est la plus importante.***

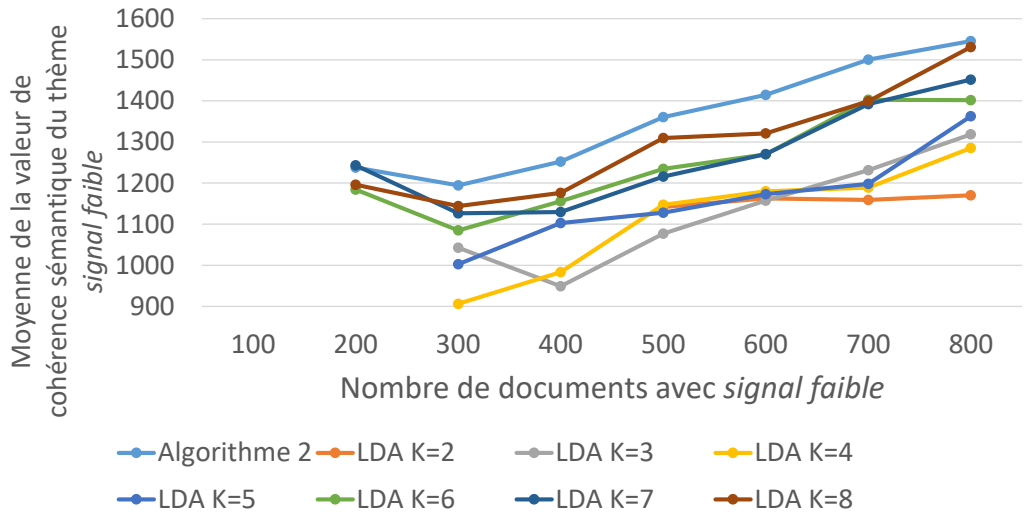


FIGURE *2.24* – ***Moyenne sur 10 tests de la valeur de cohérence sémantique du thème*** signal faible ***"détecté" avec l'algorithme 2 comparée aux 7*** LDAs ***originaux paramétrés avec*** k ***variant de*** 2 ***à*** 8. ***L'algorithme 2 détecte le thème*** signal faible ***avec la plus grande valeur de cohérence comparée à celles de*** LDA.

2.7.2.2 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 (cf. Exemple 2.25), chacun relatif à une pathologie parmi 23 présentes dans le corpus [HBLH94].

> Effect on neutrophil kinetics and serum opsonic capacity of intravenous administration of immune globulin to neonates with clinical signs of early-onset sepsis
> This study was designed to test the hypothesis that administration of immune globulin to human neonates with early-onset bacterial sepsis would (1) facilitate neutrophil egress from the marrow, (2) improve serum opsonic capacity, and (3) facilitate recovery from the infectious illness...

FIGURE *2.25* – ***Un échantillon de document original provenant du corpus Ohsumed.***

Nous proposons d'ajouter artificiellement un(e) 24ème pathologie/Thème 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 (cf. Exemple 2.26). Le jeu de données supplémentaire est accessible en suivant ce lien : `https://doi.org/10.5281/zenodo.3591580` [Mai19b]

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

Reston virus
Reston virus (RESTV) is one of six known viruses within the genus Ebolavirus. Reston virus causes Ebola virus disease in non-human primates ; unlike the other five ebolaviruses, it is not known to cause disease in humans, but has caused asymptomatic infections. Reston virus was first described in 1990 as a new "strain" of Ebola virus (EBOV). It is the single member of the species Reston ebolavirus, which is included into the genus Ebolavirus, family Filoviridae, order Mononegavirales. Reston virus is named after Reston, Virginia, US, where the virus was first discovered...

FIGURE *2.26* – ***Un échantillon de document original provenant de Wikipedia et traitant de maladies connexes à Ebola.***

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 Word2vec-on-wikipedia [7]. Nos tests utilisent la méthode de pondération *tf-idf* pour choisir les mots sur lesquels calculer la cohérence des thèmes et construire l'arborescence des thèmes (à la place de la pondération obtenue par *LDA*) (cf. Equation 2.5). Compte tenu des spécificités de ces documents ayant attrait au domaine médical, nous choisissons d'utiliser les poids *LDA* pour révéler la sémantique du thème et les valeurs de *tf-idf* pour détecter les mots-clés saillants du thème.

$$\textit{tf-idf}_i = f_{t_i,D_i} \cdot \log \frac{|D|}{|D_i|} \quad \text{où} \quad f_{t_i,D_i} = \frac{1}{|D_i|}\left(\sum_{j=0}^{D_i} \frac{n_{i,j}}{n_j}\right) \quad \text{et} \quad D_i = \{d_j : t_i \in d_j\} \tag{2.5}$$

$\textit{tf-idf}_i$ 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_i le nombre de documents où le terme t_i apparaît et $n_{i,j}$ le nombre d'occurrence de t_i dans le document d_j. f_{t_i,D_i} représente le calcul de fréquence d'un mot que nous avons choisi. Il correspond à la moyenne du nombre d'apparition du mot t_i dans tous les documents où il apparait D_i (Remarque : il s'agit d'une version légèrement modifiée pour prendre en compte la diversité des tailles de documents)

Une fois cette pondération appliquée sur l'ensemble des mots de chaque thème, ceux-ci sont triés par ordre décroissant de cette valeur. La cohérence de chaque thème 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 thèmes.

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

Nom du thème	thème 1		thème 2		thème 3		
Cohérence du thème (I_1)	32.59		29.94		25.29		
LDA $k =$	15		25		20		
Premiers mots du thème	scurfy	35.66	paroxetine	52.32	ebola	150.22	
	pcnsl	35.64	ntds	47.77	chikungunya	134.07	
	spm	35.64	gbp	38.05	song	109.39	
	dracunculiasis	33.41	dexmedetomidine	35.64	deworming	90.36	
	coc	31.18	mor	35.64	announces	85.61	
	aerd	29.95	mefenamic	33.29	soil-transmitted	57.07	
	agep	29.95	deslorelin	33.29	vinson	55.69	
	alkaloidal	29.95	alfuzosin	33.29	ebolavirus	55.31	
	agn	27.00	btb	33.29	ntd	52.32	
	rosacea	25.96	dabao	28.54	ntds	47.77	

thème 4			thème 5		...
21.03			19.94		...
20			20		...
	mvb	53.46	ngc	66.58	
	pni	52.32	oxy	60.14	
	litho	47.56	efamol	53.46	
	qsp	38.51	cmh	52.32	
	pvri	38.05	fet	49.00	
	rsi	38.05	nrv	42.80	...
	vge	35.64	tpcs	42.80	
	duncan	35.06	phr	42.80	
	nmd	33.29	parvalbumin	42.80	
	impactor	33.29	mcj	38.05	

TABLE *2.5* – ***Expérimentation sur un corpus de comptes-rendus médicaux avec ajout d'un*** signal faible ***sous forme de documents portant sur des agents infectieux. Présentation des résultats obtenus sur les*** 5 ***premiers thèmes (parmi les*** 13 ***thèmes triés selon leurs cohérences) obtenus après construction de l'arbre et élagage. Pour chaque thème, est indiquée 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.

Le tableau 2.5 montre les résultats obtenus sur le corpus Ohsumed modifié. Après élagage, 13 thèmes sont retenus. On remarque que les mots relatifs aux documents portant sur les maladies infectieuses tropicales (*signal faible*) sont capturés par un thème classé 3ème / 13 en terme de cohérence. Les mots courts semblent inintelligibles mais il s'agit d'acronymes médicaux. Le fait que le thème *signal faible* ne possède pas la cohérence la plus élevée (mais seulement la 3ème sur 13) parmi les thèmes détectés de l'arbre (après élagage) provient de la nature même des documents injectés. Ils 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 2.27 montre les 10 premiers mots de l'ensemble du corpus triés selon leur valeur

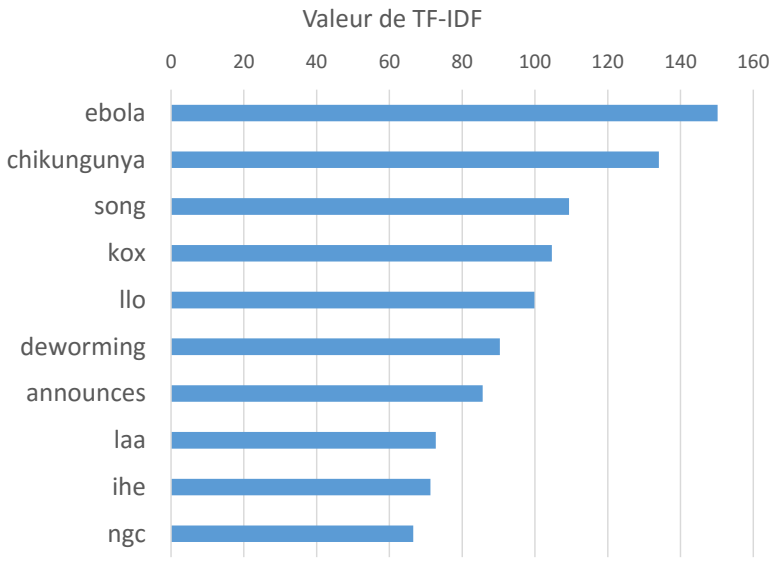


FIGURE *2.27* – ***Liste des mots du corpus ayant les plus grandes valeurs de pondération*** tf-idf***. On remarque que les mots clés du thème*** signal faible ***détecté sont parmi les premiers de la liste.***

de pondération *tf-idf*. Parmi les 10 premiers mots, 5 appartiennent au Thème *signal faible* : “ebola”, “chikungunya”, “song”, “deworming” et “announces”. Les résultats montrent la pertinence de notre solution (indicateur et algorithme) dans la recherche de *signaux faibles* dans ce contexte de corpus médical augmenté de documents ayant trait à des maladies infectieuses tropicales.

L’exploitation combinée des thèmes obtenus par l’algorithme 2 avec la liste des mots du corpus triés par leur valeur de pondération *tf-idf* permet la détection de mots portant le thème *signal faible*. Ceux-ci connus, il est alors possible d’enrichir cette liste par des phases d’exploration sur le réseau (cf. chapitre 3).

2.8 Conclusion

Pour l’analyse des documents et l’extraction du *signal faible*, nous adoptons une modélisation thématique multi-niveaux par l’approche conjointe *LDA/Word2Vec*. Nous appliquons *LDA* sur l’ensemble des documents, tout en faisant varier le nombre de thèmes désirés, 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 thèmes où au moins l’un d’entre-eux est susceptible de contenir les mots-clés d’un *signal faible* potentiel. Pour détecter *in fine* ce dernier, la méthode de pondération *tf-idf* est calculée pour chaque mot afin de détecter les mots porteurs du thème *signal faible*.

Les différents tests montrent que la modélisation thématique multi-niveaux conduit à la sélection des thèmes cohérents au sens de la définition du *signal faible* ainsi qu’à la détection

du thème ou de plusieurs thèmes porteurs de *signaux faibles* potentiels lorsque ce critère est appliqué uniquement sur les mots rares (mots non-communs). L'approche *LDA* conditionne la recherche de thèmes situés sur le même niveau d'arborescence (k fixé *a priori*) : certains thèmes peuvent être cohérents et d'autres non à ce niveau de décomposition. La recherche des thèmes les plus pertinents nécessite une analyse approfondie de l'arbre de partitionnement, ce qui est effectuée par l'algorithme 2.

La modélisation thématique multi-niveaux répond aux caractéristiques retenues du *signal faible* :

— le *signal faible* est caractérisé par un faible nombre de mots par document et présent 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 relié), 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 thème les associations très locales. L'élagage permet d'isoler des thèmes, 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*).

Chapitre 3

Agent mining et développement d'un logiciel

Dans le chapitre précédent, notre approche de modélisation thématique multi-niveaux proposée a permis d'obtenir des thèmes où au moins l'un d'entre-eux est susceptible de contenir les mots-clés d'un *signal faible* potentiel.

Ce chapitre présente la seconde partie de la solution de notre chaine de traitement de données. Celle-ci propose une solution d'*agent mining*, combinaison d'algorithmes issus du *data mining* et des systèmes multi-agents. Trois systèmes multi-agents sont mis en place, un premier aura pour tâche la recherche d'information, le deuxième, respectivement le troisième, est utilisé pour l'interaction des documents, respectivement des mots, avec l'utilisateur. Au sujet du deuxième, des agents, représentant des documents, animés par des forces d'attraction/répulsion, se déplacent dans un espace 3D. L'utilisateur peut formuler des requêtes, figeant un ou plusieurs agents réorganisant ainsi l'affichage des autres agents autour des agents immobiles. Les agents de recherche enrichissent le corpus de nouveaux documents par des requêtes sur des moteurs de recherche.

Dans ce chapitre, nous présentons un état de l'art pour les disciplines que sont le *data mining*, les Systèmes Multi-Agents ainsi que l'*agent mining*, combinaison des deux disciplines. Nous décrivons les choix retenus dans notre solution pour répondre à l'étude dans le temps des thèmes découverts par l'approche de modélisation thématique multi-niveaux. Nous détaillons ensuite comment l'ensemble de la solution est mise en œuvre dans un logiciel et expliquons les choix réalisés dans la conception.

Sommaire

2.1 Introduction 62
2.2 Modèle thématique 63
2.2.1 L'analyse sémantique latente 63
2.2.2 L'analyse sémantique latente probabiliste 64
2.2.3 L'allocation de Dirichlet latente 65
2.3 Word embedding 66
2.4 Justification de l'approche *LDA*... 68
2.5 Justification d'une approche conjointe... 70
2.6 *LDA* augmenté avec *Word2Vec* 71
2.6.1 Cas d'utilisation de *LDA* sur Wikipédia 73
2.6.2 Indicateur de cohérence en tant que mesure intra-thème 75
2.6.3 Recherche du paramètre k conduisant aux thèmes les plus pertinents au sens du critère de cohérence 76
2.6.4 Une approche heuristique pour déterminer les thèmes les plus pertinents sur l'ensemble de l'arborescence *LDA* 77
2.7 Expérimentations 79
2.7.1 Tests sur un corpus artificiel 80
2.7.2 Tests sur des corpus de données réelles 89
2.8 Conclusion 97

3.1 Introduction

Le signal partiel obtenu, la connexion à un contexte d'information plus large pour poursuivre les investigations en ayant recours à d'autres médias, doit permettre d'évaluer les potentiels corrélations et les enjeux auprès de décideurs. Des méthodes d'extraction de connaissances, comme le Web Mining, et les systèmes multi-agents peuvent fournir des solutions d'enrichissement de l'information et de présentation des résultats sous forme d'interface dynamique.

Dans ce chapitre, nous nous intéressons au problème de la représentation des documents dans un espace 3D, et au suivi temporel de *signaux faibles* potentiels et de leur connexion à un contexte plus large par de la fouille de données sur le Web. Pour cela, le domaine de l'*agent mining*, apporte des solutions par l'intégration et l'interaction entre *data mining* et Système Multi-Agents.

Nous commençons par décrire le domaine de l'extraction de connaissances (ou fouille de données) suivi des techniques de fouille du Web. Nous présentons également le domaine des systèmes multi-agents appliqués aux documents. Nous continuons par la mise en œuvre de

deux systèmes multi-agents appliqués à la fois aux documents et aux mots pour l'identification de *signaux faibles*. Enfin, nous présentons les fonctionnalités du logiciel développé qui intègre l'ensemble de la chaine de traitement et met en œuvre des services d'interaction.

3.2 Etat de l'art

3.2.1 Extraction de connaissances

Les données produites en grande quantité dans le monde chaque jour font émerger des besoins d'extraction d'information toujours plus pertinents. Les réseaux sociaux, les bases de données Open-Source gouvernementales sont autant de sources avec lesquelles les chaines de traitement doivent composer pour corréler de l'information utile. Les méthodes d'extraction de connaissances permettent de traiter les données non-structurées avec des processus d'exploration, d'analyse, de modélisation et de visualisation. L'information extraite peut prendre diverses formes (résumé, graphique, modèle, ...) et doit permettre la prise de décision [FPSS96, PKS15, HV17]. Ces nouvelles techniques doivent être toujours plus performantes pour traiter le volume grandissant de données générées.

Le problème qu'adressent les outils d'extraction de connaissances est la recherche de sens par l'extraction d'informations, à partir de données faiblement structurées, concises, compréhensibles, valides et utiles, ces dernières pouvant se présenter sous plusieurs formes, comme des synthèses ou des modèles de prédictions.

Le processus d'extraction de connaissances (cf. Figure 3.1) est un processus itératif et nécessite des rétroactions afin d'ajuster chacune de ces étapes. Ce processus ne peut être décrit par de simples relations mathématiques car il nécessite sans cesse des ajustements selon les données utilisées et les résultats que l'on souhaite produire [HKP12].

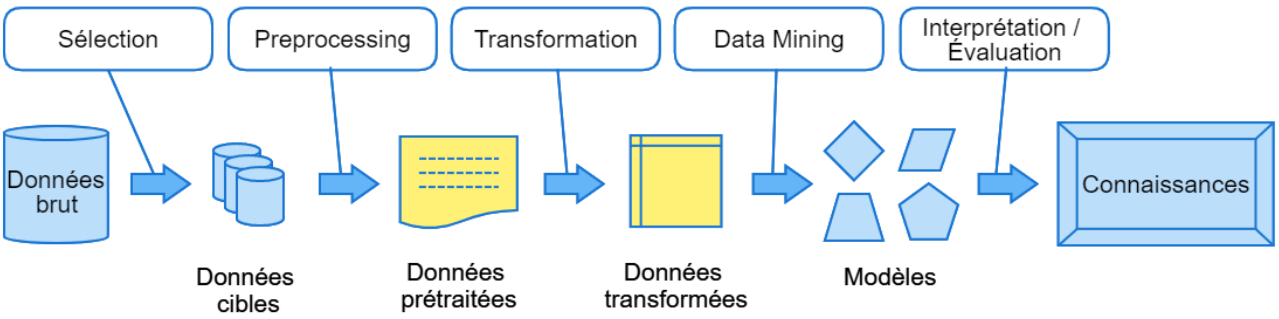


FIGURE *3.1* – ***Le processus d'extraction de connaissances***

Ce processus suit généralement les sept étapes suivantes [MR09] :

En amont de l'application de la chaine de traitement. Il est nécessaire d'avoir une compréhension du domaine sous analyse, de définir les objectifs et résultats visés, et de préciser alors les différentes étapes du traitement à mettre en œuvre. Il faut définir sous quelle forme l'information sera produite et celle qui est déjà connue du système.
Sélection. La sélection des données qui seront utilisées est réalisée à cette étape. Les données doivent être regroupées, et de nouvelles peuvent être nécessaires pour disposer d'une base de données la plus descriptive et complète possible. Le choix des paramètres de traitement est alors défini pour s'adapter aux caractéristiques des données.
Preprocessing. Cette étape consiste à nettoyer les données afin de corriger les informations incomplètes ou de supprimer les informations non utiles aux résultats, ainsi que le bruit entachant les données.
Transformation. Des méthodes sont utilisées pour préparer les données à l'étape suivante du *data mining* (DM). Il est possible par exemple de réduire la dimension ou d'effectuer des transformations de paramètres. Celles-ci dépendent des outils utilisés et des données traitées.
Data mining. Dans cette étape, nous choisissons la méthode de DM appropriée en fonction de l'information recherchée. Le modèle choisi (par exemple clustering, classification, régression, ...) est fonction des étapes précédentes.
Le modèle défini est alors appliqué sur les données transformées. Il peut être nécessaire d'appliquer plusieurs fois cette étape dans le but d'obtenir les résultats souhaités et en fonction des paramètres de l'algorithme choisi.
Interprétation/Evaluation. Durant cette étape, la qualité des résultats est évaluée au regard notamment de ce qui était attendu à l'étape de **Sélection**. Une analyse de la compréhension et de la pertinence des résultats est réalisée pour documenter et évaluer la chaine de traitement.
Post-Exploitation. Les résultats obtenus servent à de futures actions. La méthode, testée en laboratoire sur des données synthétiques, est alors mise en œuvre sur des cas réels proposant des données dynamiques, nécessitant un ajustement des paramètres souvent crucial.

3.2.1.1 *Data mining*

Le *data mining* est une étape du processus d'extraction de connaissances ("ECD" pour "Extraction de Connaissances à partir des Données" ou "KDD" en anglais) qui se compose d'algorithmes et méthodes de fouilles de données afin de révéler l'information présente dans un important volume de données pré-traitées. Cette étape se réfert à l'analyse de données. Ces méthodes et techniques sont issues de plusieurs domaines comme les statistiques, les mathématiques ou l'informatique [MR09].

Les techniques de *data mining* ont déjà été appliquées à un grand nombre de domaines applicatifs [HKP12, hM14, Bra16]. On peut citer dans les domaines des signaux et images, l'analyse de données satellitaires, la prévision de la consommation électrique et les diagnostics médicaux. Dans le domaine du document, on retrouve la détection de fraude et la synthèse

de texte. Pour des domaines multi-systèmes complexes, il y a par exemple l'optimisation de centrales thermiques, l'analyse de risque toxicologique et la prévision météorologique.

Ces données extraites peuvent servir à la prise de décision si la qualité des résultats s'avère pertinente pour l'utilisateur. Dans le cas contraire, l'utilisateur peut choisir de changer les paramètres/hyperparamètres de la méthode ou changer de méthode s'il cherche d'autres informations.

On peux référencer six grandes catégories de méthodes utilisant deux types de données différentes [FPSS96, HKP12, hM14, Bra16] :

- **Apprentissage supervisé**. Après une phase de labellisation, et éventuellement d'apprentissage, l'objectif est de prévoir des valeurs d'attributs dans des cas non-vus sur la base de données fournie pré-annotée. On dit que les données sont "labellisées".
  - **Classification**. La classification est une méthode très utilisée pour le *data mining*. Elle consiste simplement au tri des données selon différentes approches comme la recherche de voisins proches, des règles de classification ou par arbre de classification.
  - **Régression**. La régression permet d'estimer les relations entre une variable dépendante, et une ou plusieurs variables indépendantes par l'apprentissage d'une fonction. La forme courante de régression est la régression linéaire permettant d'estimer une relation linéaire sur des données en fonction d'un critère mathématique spécifique. Cette relation fait correspondre un élément de données à une ou plusieurs variables de prédiction explicative.
  - **Détection d'anomalies (*anomaly/outlier/deviation/change detection*)**. La détection d'anomalies se concentre sur la recherche et l'identification d'éléments, d'événements ou d'observations rares par rapport aux valeurs mesurées ou normatives et qui soulèvent des suspicions en s'écartant de la majorité des données. Ces données, considérées comme aberrantes ou anormales, peuvent être détectées aux moyens de tests statistiques via des modèles de distribution ou de probabilité, ou à l'aide de mesures de distance. D'autres méthodes existent comme les méthodes s'appuyant sur la densité afin d'identifier les valeurs anormales dans une région locale qui sembleraient cependant normales lorsqu'on s'appuie sur une distribution statistique globale.
- **Apprentissage non supervisé**. Les données n'ayant pas d'attribut désigné sont appelées données non "labellisées". Les techniques de DM relèvent dans ce cas des méthodes d'apprentissage non supervisées et visent à extraire de l'information sur les données fournies.
  - **Groupement par similitude / Règles d'associations *association rule learning / dependency modeling***). Ces techniques recherchent des relations entre des valeurs de variables pour former des règles d'associations. Les règles peuvent être nombreuses et dépendantes des données. Leurs fiabilités sont indiquées par une probabilité définissant la relation entre plusieurs variables selon leurs valeurs.

- **Regroupement (*Clustering*)**. Le clustering rassemble des individus en groupe selon des similarités dans les valeurs de leurs attributs. Les individus d'un même groupe partagent de forte similarité entre eux mais diffèrent des individus des autres clusters.
- **Résumé automatique de texte (summarization / feature extraction)**. Souvent appliquée à l'analyse exploratoire interactive des données et à la génération automatisée de rapports, la synthèse met en œuvre des méthodes dans le but de trouver une description compacte d'un sous-ensemble de données qui représente les informations les plus importantes ou les plus pertinentes au sein du contenu original. Il existe 3 méthodes principales pour générer des résumés de texte :
  - **Abstraction**. L'approche construit une représentation sémantique interne du contenu, puis l'utilise pour générer des phrases créant ainsi un résumé, pouvant transformer une partie du contenu extrait en paraphrasant des sections du document source. Cette transformation s'avère plus complexe à mettre en œuvre que les autres méthodes car elle nécessite des traitements du langage naturel et une compréhension approfondie du domaine du texte original.
  - **Extraction**. Cette approche extrait des phrases complètes censées être pertinentes dans le document et les concaténe pour produire un extrait.
  - **Compression**. Les phrases extraites sont compressées afin d'éliminer l'information superflue.

3.2.1.2 Web Mining

Le web est la plus grande base de données dans le monde constament en expansion en volume et dimension de données. Ce contenu est consommé par des utilisateurs pour des besoins variés. Les techniques de *data mining* appliquées au Web sont appelées Web Mining (Fouille du Web) [DC12, HKD13]. Ces dernières recherchent de l'information et de la connaissance provenant de documents Web, des liens entre les documents, des logs d'usage de sites webs,...

Le web offre un grand nombre de challenges et d'opportunités nécessitant d'adapter des approches classiques du *data mining* à la complexité de la fouille de données sur le web et les réseaux sociaux : grand volume de données, hétérogénéité, diversité, fortement dynamique, ponctuelle ou s'étendant sur le temps, très dépendante de la source (en termes notamment de fiabilités), redondance, à différents niveaux et d'échelle de visibilité,... (cf. Tableau 3.1)

Les techniques de Web Mining peuvent être classées selon trois types d'approche afin d'extraire de la connaissance [HKD13, JP17] :

- **Fouille du contenu Web (Web content mining)** (cf. Figure 3.2). Son objectif est d'extraire les informations utiles, souvent de type multimédia, présentes sous forme de ressources web comme textes, images, sons et vidéos. Ces techniques doivent identifier et extraire ces informations multimodales. Elles s'appliquent généralement à des sources

Caractéristiques du Web				
Semi-structuré, évolutif, hétérogène, redondant, bruité, traite tous les domaines				
Forme	**Echelle d'interaction**	**Représentation**	**Contenu**	**Accessibilité**
Données	Personnes	Images	Information principale	Surface (pages web)
Informations	Organisations	Textes	Publicité	Profondeur (base de données)
Services	Systèmes automatisés	Tableaux	Liens de navigation	
		Données multimédia	Information de droits d'auteurs	

TABLE *3.1* – ***Description des caractéristiques du Web [LC04, DC12, HKD13]***

de données hétérogènes (documents HTML, réponses à des requêtes sur des bases de données ou bibliothèques numériques) et s'apparentent aux techniques de recherche d'information ciblée (Information Retrieval). Des données, comme le prix ou la description d'un produit, des messages portant un certain caractère dans des forums, peuvent être extraites par ces méthodes. Des techniques plus avancées permettent d'extraire des sentiments d'avis clients et d'échanger sur des forums. Celles-ci ne font pas partie des tâches classiques de *data mining*. Un grand nombre de techniques de fouille de contenus utilisent des méthodes d'extraction de texte (Text Mining) car la plupart des données se présentent sous cette forme [DC12, HKD13].

Les approches de fouille de contenus du Web sont très dépendantes de l'application du fait de leur complexité créative et par la nature semi-structurée des ressources, *a contrario* des méthodes de Text Mining qui se concentrent spécifiquement, le plus souvent sur les textes non structurés. Elles forment un domaine de recherche ouvert multidisciplinaire en constante évolution et tiré par les besoins des décideurs avides d'outils toujours plus performants pour extraire l'information utile.

Dans le cadre de notre seconde étape dans la chaine de traitement, nous mettons en œuvre une solution de fouille du contenu. Celle-ci permet, à partir des résultats obtenus par des moteurs de recherche, de récupérer de nouveaux documents afin d'enrichir l'information déjà obtenue portée par les *signaux faibles* potentiels. Des agents dédiés à cette tâche, appelés agents de recherche, récupèrent le contenu pertinent dans des pages Web et créent de nouveaux agents avec ces données.

La solution que nous proposons est adaptable à différentes sources de données. Le service de recherche Web permet actuellement la récupération de documents sur les moteurs de recherche Qwant/Web et Elastic. Il peut-être étendu à d'autres sources de données (bases de données, sites spécialisés,. . .).

— **Fouille de la structure du Web (Web structure mining)** (cf. Figure 3.3) consiste en la découverte et l'analyse de la structure d'ensemble ou d'un sous-ensemble de pages

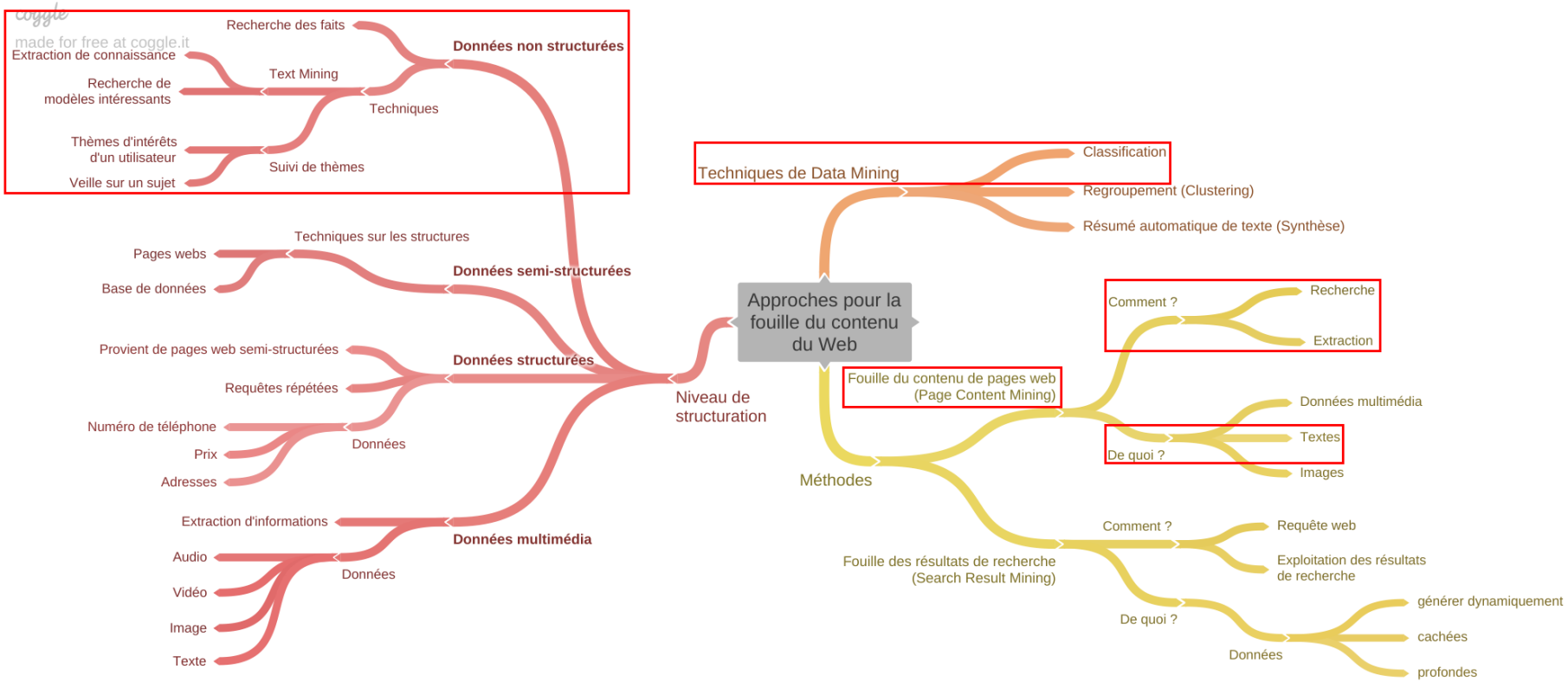


FIGURE *3.2* – ***Carte heuristique des approches pour de la fouille du contenu Web. Nous détaillons les différents niveaux de structuration et formes de données du Web. On peut identifier deux groupes de méthodes parmi les techniques de fouille du contenu du Web : "Fouille du contenu de pages web" et "Fouille de résultats de recherche". Les techniques principalement employées sont la classification, le regroupement ainsi que le résumé automatique de texte. Dans notre étude, nous nous positionnons selon les cadres rouges. Les agents de recherche extraient de l'information à partir de données non structurées. Celles-ci permettent de suivre des signaux faibles potentiels. A partir de résultats sur des moteurs de recherche, les agents de recherche fouillent le contenu de pages Web et créent de nouveaux agents. Ces agents se classifient dans un système multi-agents à partir de l'environnement de ce dernier comprenant les agents déjà présents.***

web via notamment l'organisation des hyperliens qui les relient. La connaissance utile est découverte à partir de la topologie de la structure des liens entre les pages définissant la structure du site. A partir de ces informations, il est possible de découvrir aussi bien les pages importantes ou les plus pertinentes que des communautés d'utilisateurs qui partagent des intérêts communs sur des réseaux sociaux. L'objectif, dans ce dernier cas, est d'améliorer les résultats de la recherche pour les utilisateurs. Les techniques de *data mining* généralistes ne permettent pas d'exploiter de telles informations car elles n'étudient pas par exemple les structures dans un tableau relationnel de base de données. Dans le cas d'un site web, celui-ci est représenté sous la forme d'un graphe dont les pages forment les noeuds, et les hyperliens définissent les arrêtes connectant les pages. Les hyperliens sont nommés in-links quand ils proviennent d'une autre page et pointent vers le document web cible et out-links dans le sens inverse. Dans l'analyse des liens, on utilise les termes "degré entrant" pour le nombre d'hyperliens pointant un noeud particulier et "degré sortant" pour le nombre d'hyperliens d'un noeud particulier vers d'autres noeuds.

— **Fouille de l'usage du Web (Web usage mining)** (cf. Figure 3.4) connue aussi sous le nom de fouille de journaux/logs du web (Web log mining), elle a pour principal objectif la recherche d'information d'utilisation et de comportement de l'utilisateur lors de la navigation sur des sites web : mieux comprendre son comportement de navigation pour mieux répondre aux besoins. La fouille du contenu ou de la structure du web utilise des données primaires du web (pages webs, contenu multimédia,...). A contrario, la fouille de l'usage utilise les données d'activités de l'utilisateur grâce au serveur web qui enregistre les derniers journaux du serveur. Ceux-ci représentent une grande quantité d'information à analyser pour étudier la valeur individuelle d'un client ou étudier l'impact d'une campagne promotionnelle.

3.2.2 Système Multi-Agents

Les systèmes multi-agents (SMA) proposent des approches multi-composants autonomes pouvant interagir entre eux et avec leur environnement pour la résolution de problèmes complexes. Le nombre vaste de disciplines utilisant les SMA montre la richesse de ce domaine de recherche. Il existe une grande multiplicité de modèles d'agents, d'environnements, d'interactions et d'organisations possibles [DGB+05, hM14, PKS15, HV17, DKJ18].

Un système multi-agents est un environnement où un ensemble d'entités autonomes interagissent selon certaines relations. Chaque agent forme une entité caractérisée, autonome et peut être de différentes natures comme un processus, un robot, un être humain, etc. Les agents résolvent des tâches en collaborant, offrant plus de flexibilité de par leur capacité à apprendre et prendre des décisions autonomes. Les agents utilisent leurs connaissances pour décider les actions à entreprendre dans l'environnement afin de résoudre les tâches qui leur ont été attribuées.

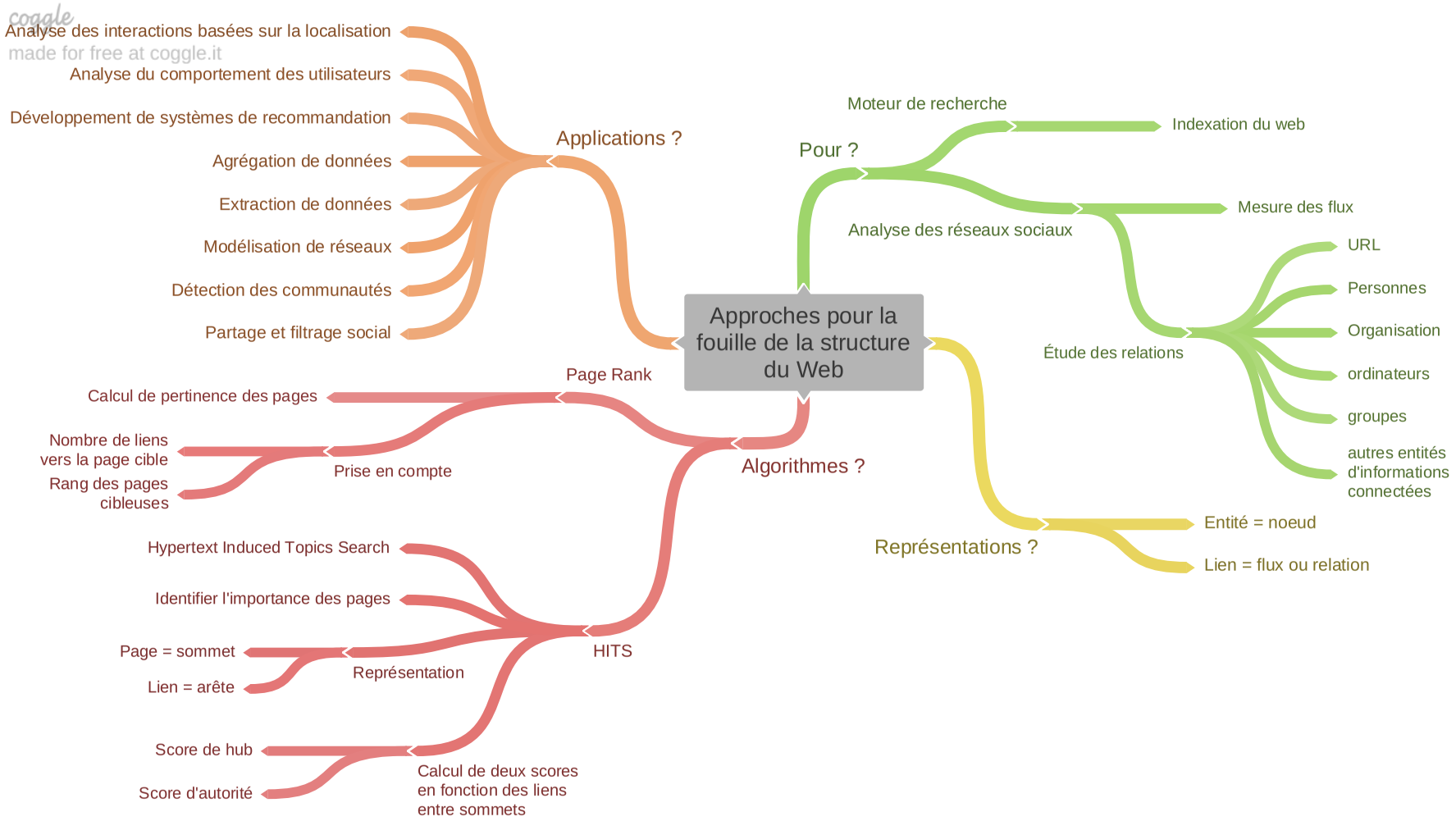


FIGURE *3.3* – ***Carte heuristique des approches de la fouille de la structure du Web. Ce schéma détaille les différentes applications, algorithmes, utilisations et représentations dans les approches d'extraction d'information sur la structure du web. Deux algorithmes largement employés existent : "PageRank" et "HITS". Les approches, pour la fouille de la structure du Web, permettent un nombre important d'applications et sont principalement employées dans des moteurs de recherche ainsi que pour l'analyse de réseaux sociaux. La représentation couramment utilisée associe les entités à des noeuds et les flux ou relation à des liens.***

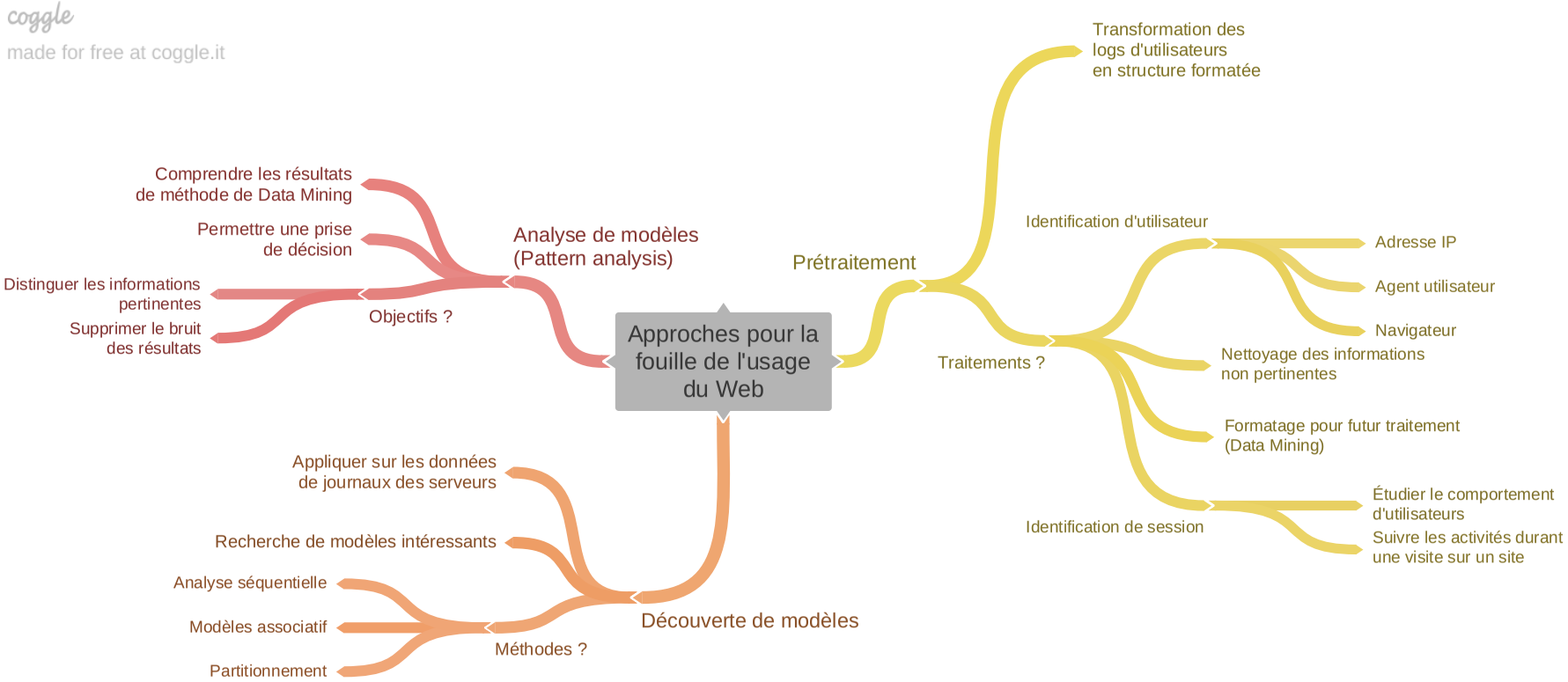


FIGURE *3.4* – ***Carte heuristique des approches pour la fouille de l'usage du Web. Dans la fouille de jounaux du web, trois phases principales sont utilisées. Le prétraitement, rassemblant un ensemble de techniques allant du nettoyage des données à l'identification des utilisateurs, transforme les logs d'utilisateurs en structure formatée pour de futur traitement avec des méthodes de*** data mining***. La "découverte de modèles" recherche des modèles intéressants à appliquer sur des données de journaux de serveurs au moyen de différentes méthodes. La dernière phase, l'analyse de modèle, appliquée après les méthodes de*** data mining ***doit permettre de comprendre les résultats pour une prise de décision.***

Ces connaissances se construisent via les interactions avec d'autres agents et l'environnement.

Les caractéristiques des SMA leur permettent d'être adaptées à de nombreux problèmes dans diverses disciplines, notamment l'informatique, le génie civil ou le génie électrique. Mais certains défis doivent encore être résolus comme la coordination des agents, l'apprentissage, la sécurité,... [RVR10]

De part leur nature, les SMA facilitent la décomposition d'un problème complexe en un ensemble de petits problèmes plus simples à résoudre créant ainsi un avantage collectif attribuable à l'interaction entre agents. Les systèmes multi-agents peuvent être de nature distribuée et dont les données sont décentralisées. Dans les SMA distribués, les notions de sécurité et de protocoles d'échanges sont des problèmes cruciaux.

3.2.2.1 Définition d'un agent

Un agent est un logiciel utilisé pour réaliser une tâche spécifique à sa conception au nom d'une autre entité comme un individu ou un autre programme. Il existe plusieurs définitions pour un agent selon diverses caractéristiques spécifiques d'applications [DGB+05, PKS15, HV17].

La définition pour Russell et Norvig est "Un agent intelligent est défini comme tout ce qui peut être considéré comme percevant son environnement par des capteurs et agissant sur cet environnement par des actionneurs" [RN10, PKS15]

Jennings et Wooldrige spécifient qu'un "agent est un processus informatique autonome qui est situé dans un environnement avec lequel il va interagir pour lui permettre d'atteindre les objectifs fixés lors de sa conception. Un agent perçoit son environnement par le biais de capteurs qui lui sont connectés" [JW98, PKS15].

La notion d'agent est générique car applicable dans de nombreuses disciplines. Parmi les définitions, on peut distinguer des caractéristiques fondamentales des agents. Un agent est une entité flexible car il doit être [hM14, PKS15] :

- **Autonome**. Un agent agit sans influence d'autres agents ou humain et contrôle ses propres actions ainsi que son état interne.
- **Proactif**. Un agent doit être opportuniste, orienté vers un objectif et prendre des initiatives plutôt que simplement agir en réponse à son environnement.
- **Réactif**. L'agent, en percevant son environnement, doit réagir au changement en réponse à un stimulus
- **Sociable**. Un agent doit être capable de savoir quand intéragir avec d'autres agents ou êtres humains afin de remplir son objectif et d'aider les autres dans leurs activités.
- **Coopération**. Un agent coordonne ses actions avec d'autres agents pour atteindre un objectif commun.
- **Mobile** : Un agent doit pouvoir se mouvoir dans l'environnement

— **Adaptable**. Il doit pouvoir évoluer et s'adapter au changement dans son environnement. Il peut changer de comportement.
— **Rationnel**. Un agent œuvre pour atteindre son objectif et n'agit pas de manière à l'empêcher d'atteindre son but.

Un agent peut être utilisé pour automatiser un nombre varié de tâches comme la sélection de données, le nettoyage des données, le pré-traitement, la classification, le regroupement et la représentation des connaissances. Les connaissances des agents dépendent de la théorie du domaine existant. Ces connaissances guident leurs raisonnements et leurs actions dans l'environnement [DKJ18].

L'analyse évolutive des données permet de détecter des modèles cachés, construire des modèles prédictifs et identifier des valeurs aberrantes, par une exploration avancée des données. Pour les systèmes multi-agents, la connaissance est collective, sous forme de ressources et répartie entre les agents.

3.2.2.2 Catégories d'agents

Parmi tous les types d'agents observables, on peut classer les agents selon deux grandes catégories [PKS15, HV17] :

Les agents réactifs

Les agents réactifs fonctionnent avec un comportement "stimulus - réponse". L'agent a une connaissance partielle de son environnement limitée à ses perceptions et ne tient pas compte des actions passées.

Un système composé uniquement d'agents réactifs va chercher à converger vers un état décisionnel stable. Si le système y parvient, il n'est pas assuré que l'état atteint est optimal. Les systèmes uniquement réactifs contiennent un grand nombre d'agents. Le système fonctionne sur le principe de l'émergence d'un comportement "intelligent" d'un ensemble d'agents non-intelligents.

Les caractéristiques d'un agent réactif et de son système sont les suivantes :
— il dispose d'une "représentation sub-symbolique" de l'environnement, limitée à ses perceptions ;
— il ne possède pas de mémoire des actions passées et n'a pas de but défini ;
— il fonctionne sur le principe du "stimulus/réponse" ;
— il est associé à un grand nombre d'agents qui sont homogènes (ayant le même comportement) ;
— il contribue à l'objectif via le principe d'émergence et non de volonté d'organisation.

La solution que nous proposons dans ce chapitre met en œuvre des agents "document", respectivement agents "mot", représentation des documents, respectivement

des mots. Ces agents réactifs disposent d'une vision partielle de l'environnement. A chaque itération, ceux-ci calculent des forces en fonction de la dynamique du système (positionnement des autres documents). Ce système auto-organisé permet une visualisation avec des interactions simples et intuitives pour l'utilisateur.

Les agents cognitifs

Les agents cognitifs disposent d'une "représentation symbolique" du monde à partir duquel ils vont être capables de formuler des raisonnements pour planifier des actions. Les agents cognitifs communiquent, tiennent compte de leur passé et s'organisent selon un mode social d'organisation. Ce dernier est évoqué dans la section 3.2.2.3. Ces agents ont une approche "intelligente" construite sur la collaboration pour résoudre un problème, avec une perspective orientée sociologique.

Les systèmes uniquement composés d'agents cognitifs n'utilisent qu'un nombre limité d'agents. Ils nécessitent une plus grande quantité de ressources. Ces systèmes ne convergent pas plus facilement mais ils permettent de traiter des problèmes plus complexes qui peuvent nécessiter une plus grande abstraction.

Les caractéritiques impliquées sont :
— une "représentation symbolique" de l'environnement et des autres agents
— la prise en compte du passé et la connaissance d'un but explicite
— une organisation "sociale"
— la collaboration entre agents dans le but d'une résolution collective du problème

Parmi les agents cognitifs principaux, on peut trouver différents types d'agents :
— **Agent interface** : il est responsable des communications entre l'utilisateur et le système. Il inclut les tâches à effectuer par le système et retourne les résultats quand elles sont accomplies. Il est responsable des communications avec les agents.
— **Agent manager** : lors de la réception d'une tâche par l'agent interface, l'agent manager organise une planification des tâches pour résoudre le problème et founir aux différents agents le travail à effectuer. Le résultat est fourni à l'utilisateur par l'agent interface. L'agent manager est responsable de la bonne synchronisation des agents.
— **Agent donnée** : il fournit les ressources provenant de multiples sources aux agents de fouilles. Il conserve les métadonnées de toutes les sources de données.
— **Agent d'extraction de connaissances (mining agent)** : un agent d'extraction de connaissances implémente un algorithme de fouille de données. Il est à l'initiative du processus de fouille en choissisant la bonne méthode, les paramètres requis de la méthode, les données d'entrée,...
— **Agent de résultat** : après le processus, cet agent récupère le contenu en sortie de l'algorithme de l'agent de fouille. Il met en forme les données au moyen d'outils de visualisation et modèles de rapport qu'il maintient à jour.
— **L'agent de courtage** : l'agent de courtage a connaissance de tous les agents du système

ainsi que de leurs capacités. Il fournit les noms des agents pouvant répondre à une demande en fonction de celle-ci.

— **Agent requête** : l'agent requête est créé suite à la requête utilisateur. Il génère les requêtes pour le système sur la base des connaissances du système afin de répondre à la demande de l'utilisateur.

Notre solution met en œuvre des agents de recherche. Ceux-ci fonctionnant sur les principes des agents d'extraction de connaissances et agents de résultat, utilisent les services définis au sein de notre logiciel pour exécuter des requêtes composées de plusieurs mots-clés sur des moteurs de recherche et enrichir ainsi le corpus initial de nouveaux documents par ces recherches d'information complémentaire.

Il existe d'autres agents cognitifs selon le domaine de recherche des données comme des agents de pré-traitement pour préparer les données, de post-traitement, des agents mobiles qui traitent l'information sur des noeuds différents pour renvoyer les résultats à un agent maitre (systèmes multi-agents distribués).

Les agents peuvent avoir différents comportements dont les deux grands principaux sont :

Comportement téléonomique : l'agent suit un comportement intentionnel en poursuivant un but explicite.

Comportement réflexe : l'agent n'a pas de but fixé à accomplir et réagit aux perceptions de son environnement.

	Agent cognitif	**Agent réactif**
Comportement téléonomique	Agent intentionnel	Agent pulsionnel
Comportement réflexe	Agent "module"	Agent tropique

TABLE *3.2* – ***Tableau des catégories d'agent selon leur comportement et leur type.***

Les agents selon leur type et leur comportement font partie de catégories d'agents différents (c.f. Tableau 3.2) :

— **Agent intentionnel**. Dans les systèmes multi-agents orientés cognitifs, les agents sont la plupart du temps intentionnels car ils ont des buts fixés à accomplir.
— **Agent "module"**. Un agent "module" n'a pas de but précis et est utilisé pour répondre aux interrogations des autres agents du système.
— **Agent pulsionnel**. Cet agent a une mission fixée proche de la veille de variable comme par exemple suivre le niveau d'un réservoir. Il déclenche un comportement si son environnement ne permet plus d'atteindre son but.
— **Agent tropique**. Cet agent n'a comme seule fonction que de réagir à son environnement local. Cette modification de l'environnement peut être déclenchée par un évènement interne au système (i.e. agent pulsionnel) ou par l'environnement.

3.2.2.3 Organisation des agents

L'organisation, dans un système multi-agents, est construite sur un ensemble de rôles, relations, et de structures d'autorités. Ils ont une importance majeure sur les performances du système en termes de résultat tant sur le plan qualitatif que quantitatif. Tous les systèmes multi-agents possèdent une forme d'organisation, même implicite et informelle. Cette organisation régit la manière dont les agents intéragissent les uns avec les autres. Ces comportements de groupes peuvent aider des agents sophistiqués à réduire la complexité de leur raisonnement. La forme, la taille et les caractéristiques de la structure organisationnelle peuveut affecter le comportement du système pour servir un certain objectif. Plusieurs études montrent l'impact significatif de l'organisation d'un système sur les performances à court et long terme en fonction de différents critères comme le nombre d'agents, les objectifs et l'environnement [HL04].

Il n'existe pas d'organisation qui puisse répondre à l'ensemble des situations. Parfois, aucune organisation seule ne peut convenir à certaines situations spécifiques qui alors nécessitent une combinaison de structures organisationnelles (cf. Annexe C).

Dans la section suivante, nous présentons 3 systèmes multi-agents (cf. tableau 3.3). Dans ces derniers, différents agents interviennent pour fournir des solutions d'enrichissement de l'information et de présentation des résultats sous forme d'interface dynamique. Les systèmes multi-agents de documents et de mots ont des organisations proches de la coalition (cf. Annexe C) à ceci près qu'ils ne forment pas de groupe et non pas d'agent "dirigeant". Les agents "document" (cf. section 3.3.2.2) et agents "mot" (cf. section 3.3.4.1) changent d'état selon l'action de l'utilisateur. Ils représentent des données et sont régis par des forces d'attraction/répulsion. Les agents de recherche (cf. section 3.3.3.1) sont des agents d'extraction de connaissances utilisés pour requêter le Web, récupérer de nouveaux documents/mots et former de nouveaux agents avec ces derniers.

	Système multi-agents		
	Documents	**Mots**	**Recherche**
Organisation des agents	Combinaison/Coalition	Combinaison/Coalition	Aucune
Agents	Agents "document"	Agents "mot"	Agents de recherche
Catégories d'agents	Agent tropique	Agent tropique	Agent intentionnel (Agent d'extraction de connaissances et Agent de résultat)

TABLE *3.3* – ***Tableau des différents systèmes multi-agents appliqués aux documents, aux mots, à la recherche et description des différents agents présents dans les 3 systèmes.***

3.2.3 *Agent mining* : *Data mining* et Système multi-agents

Le *data mining* et les systèmes multi-agents ont longtemps été développés comme deux domaines disjoints. Ces disciplines ont fait l'objet d'évolutions constantes avec des objectifs distincts afin de répondre aux défis spécifiques de leur domaine. De nouveaux challenges tels que le "cloud computing", l'informatique comportementale ou sociale nécessitent l'intégration de ces deux domaines de recherche.

Chacune peut contribuer ainsi à la construction de [CGM09, CWY12] :

— solutions de ***data mining*** pour les **agents**, (*Data Mining-Driven Agents*). Les méthodes d'apprentissage, d'adaptation, d'évolution et d'analyse du comportement fournissent des solutions intéressantes pour des agents intelligents. Des modèles de DM sur le calcul distribué peut fournir de nouveaux modèles pour la répartition et la planification entre agents, aider à l'organisation et l'évolution des systèmes multi-agents. Il est possible d'utiliser le *data mining* pour de la prise de décision automatique. Par exemple, avec des techniques d'extractions de connaissances, il est possible d'identifier, extraire et analyser des logiques d'actions humaines afin de les encapsuler dans des agents robustes et fiables pour automatiser des prises de décision humaine.

— solutions d'**agents intelligents et système multi-agents** pour le ***data mining*** (*Agent-based Data Mining*). Les agents peuvent aider le processus de *data mining* en contribuant à la sélection, l'extraction et au traitement des données dans des environnements distribués ou exploitation parallèle de différentes sources d'informations pour de la classification, du clustering et du tri par règles d'association. La fouille de données dans des environnements distribués est largement utilisée dans la communauté comme solution d'agents intelligents pour le *data mining*.

  Les agents peuvent permettre une extraction interactive centrée sur l'utilisateur selon les besoins et objectifs fournis aux systèmes. Dans ces systèmes, les agents forment une interface entre l'utilisateur et les systèmes logiciels pour répondre aux besoins de la fouille de données. Dans des environnements mobiles, les agents vont surveiller des changements de données locaux et au moyen de modèles locaux dynamiques, fournir les informations à des systèmes centralisés par l'intermédaire d'agents coordinateurs.

3.3 Système multi-agents et *data mining* proposé

La phase *agent mining* se plaçant dans la chaine de traitement après celle de la modélisation thématique multi-niveaux, a pour objectif le suivi de *signaux faibles* potentiels par la recherche de nouveaux contenus en ligne et propose une visualisation interactive permettant de gérer en temps réel les documents porteurs de *signaux faibles* (cf. figure 3.5). Les mots-clés détectés lors de la première phase sont utilisés pour alimenter les systèmes multi-agents auto-organisés

(SMA), où des agents "document" ou "mot" sont animés par des forces d'attraction/répulsion construites sur des similitudes sémantiques (pour les documents) ou de contexte (pour les mots). Au moyen des thèmes détectés, des agents de recherche enrichissent le système de nouveaux documents par des requêtes sur des moteurs de recherche. Ces SMA sont décrits par les points suivants :

1. de nouveaux agents "document"/"mot" 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/mots et donc aussi des thèmes visibles ;
3. des interactions humaines sont possibles en forçant manuellement la position d'agents "document"/"mot" qui deviennent alors agent "requête" ;
4. les agents "requête" transmettent de l'information aux agents de recherche.

3.3.1 Chaine de traitement

Les documents sont modélisés à l'aide d'un système multi-agents. Chaque agent représente un document qui se déplace dans un environnement 3D (où les axes n'ont pas de signification pré-déterminée) en fonction de forces construites à partir du vecteur caractéristique du document et de ceux de ses voisins (attraction-répulsion). Constamment en mouvement, ils ne réagissent qu'aux stimuli extérieurs envoyés par les autres agents "document" dans l'espace. Les agents interagissent entre eux au moyen donc de forces qui peuvent être attractives ou répulsives. Celles-ci, calculées entre paire d'agents et selon la similarité de leurs vecteurs caractéristiques (attirés par les agents similaires, repoussés par les agents non similaires), définissent leurs comportements qui produisent, au niveau global, une auto-organisation de tous les documents dans l'espace 3D. Ce modèle, par sa simplicité, reproduit les comportements complexes communément présents dans les systèmes CBIR ("Content-based image retrieval") traditionnels tels que la navigation, les requêtes par texte et/ou image ou encore l'interaction avec l'utilisateur.

3.3.2 Système multi-agents associé aux documents

3.3.2.1 Principe

Dans ce système multi-agents, évoluent les agents "document". Ceux-ci sont représentés dans l'espace et sont régis par des forces d'attraction/répulsion qui les animent selon leur similitude sémantique (cf. figure 3.6). Ce système multi-agents auto-organisé permet à l'utilisateur d'interagir avec les agents en forçant leurs positions dans l'espace. L'utilisateur peut ainsi visualiser

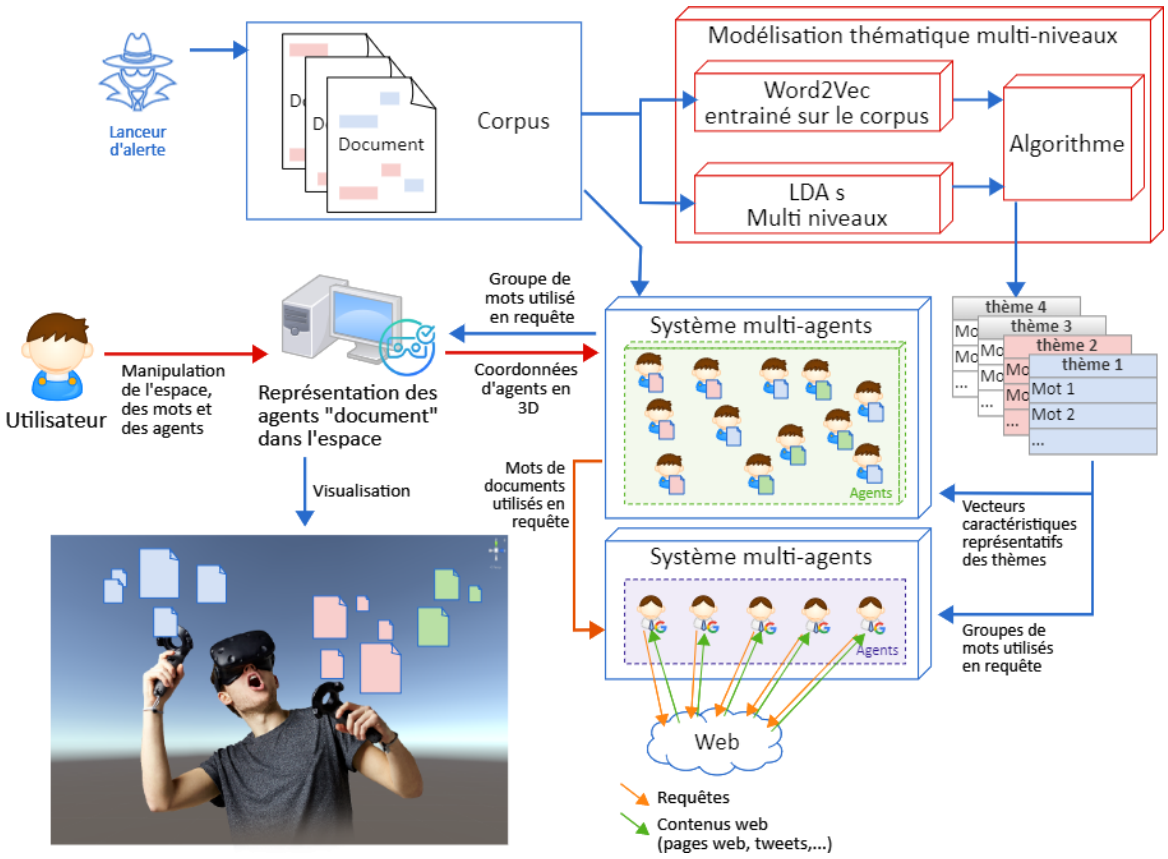


FIGURE *3.5* – ***Schéma de la chaine de traitement. Les thèmes obtenus par la modélisation thématique multi-niveaux décrite au chapitre 2 ainsi que le corpus de documents sont injectés dans les systèmes multi-agents. Une interface visuelle et interactive permet à l'utilisateur d'interagir avec le système. Ce dernier effectue alors des recherches sur le Web afin d'enrichir le corpus de nouveaux documents en rapport avec les thèmes trouvés et ceux sélectionnés par l'utilisateur. Lorsque l'utilisateur choisi de manière interactive un document, les requêtes sont construites à partir des mots-clés présents dans le document et découverts dans les thèmes. Actuellement, l'interface utilisateur permet uniquement une intéraction classique simple de type clic souris.***

le contenu des documents et étudier des documents similaires situés dans leur voisinage dans cet espace de visualisation.

La figure 3.6 montre le principe. Au moyen du système multi-agents associé à la recherche de document, la plateforme d'investigation recherche activement de nouveaux documents relatifs aux différents thèmes, augmentant progressivement la taille du corpus et découvrant d'autres mots apparentés éventuellement à ces thèmes. 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. Cette approche permet de s'ouvrir à un contexte informationnel plus large.

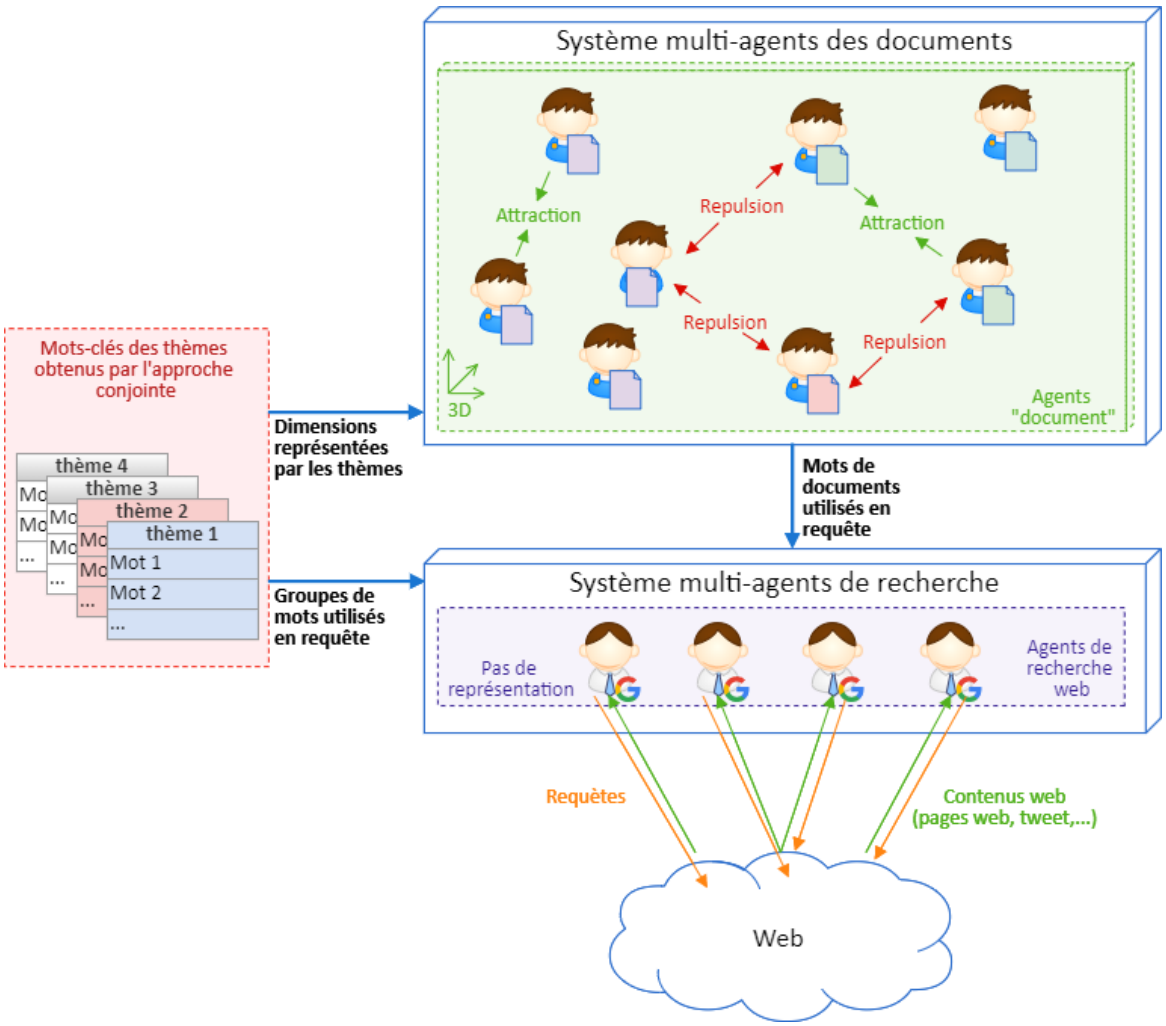


FIGURE *3.6* – ***Les agents "document" interagissent les uns avec les autres grâce à des actions de type attraction/répulsion. Les agents de recherche construisent des requêtes à partir des mots associés aux différents thèmes obtenus lors de la modélisation thématique multi-niveaux et éventuellement présents dans les documents sélectionnés par l'utilisateur (renforcement) (cf. Chapitre 2).***

3.3.2.2 Agent associé à un "document"

Chaque document est représenté par un agent évoluant dans un espace à N dimensions. Les N dimensions correspondent aux nombres de thèmes extraits de la modélisation thématique multi-niveaux (cf. Chapitre 2). On associe donc au document un vecteur caractéristique composé de N valeurs, chacune d'elles définit la composante du document associée à ce thème (cf. figure 3.7). Celle-ci est déterminée par la somme des occurences des mots, apparaissant dans le document appartenant au thème, pondérées par leur valeur de *tf-idf* :

$$C_{d_a}^{n} = \sum_{w_i \,\epsilon\, E_n \cap d_a} n_i * tf\text{-}idf_i \tag{3.1}$$

Où $C_{d_a}^n$ représente la composante du document d_a associée au thème c_n, $tf\text{-}idf_i$ la valeur de *tf-idf* du mot w_i dans le corpus de documents D, n_i le nombre d'occurrence de w_i dans d_a et E_n l'ensemble des mots retenus du thème n.

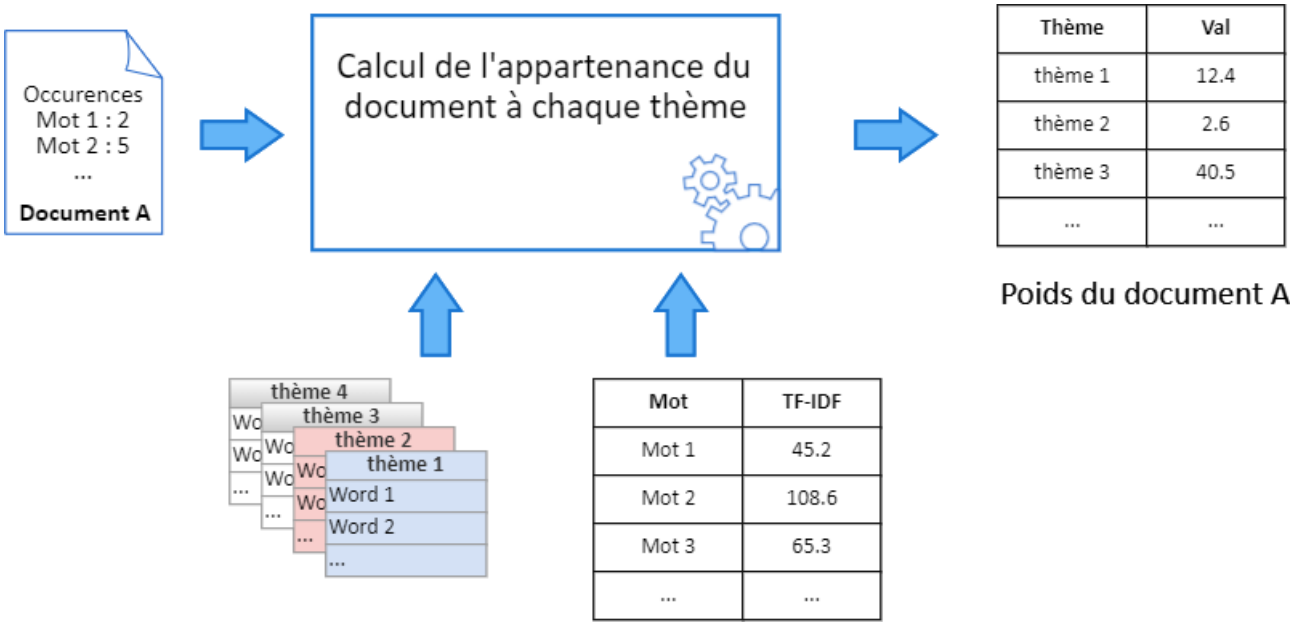


FIGURE *3.7* – ***Pour chaque document est calculé un vecteur caractéristique composé de n valeurs, chacune d'elles définit la composante du document associée à un thème. Ce vecteur caractéristique permet l'application des forces d'attraction/-répulsion dans le système multi-agents construit. Pour chaque thème, et chaque document, nous calculons le nombre d'occurence des mots du thème présents dans le document en prenant en compte leur rareté et le fait que ces mots soient présents dans peu de documents (grâce à leur valeur de tf-idf).***

Chaque document est défini par ces composantes associées à chaque thème et représentées par un vecteur caractéristique dans l'espace des thèmes. Les vecteurs caractéristiques sont utilisés pour calculer les forces d'attraction/répulsion entre pairs d'agents "document".

3.3.2.3 Attraction/répulsion entre pairs d'agents "document"

La similarité, respectivement dissimilarité, des agents est représentée par une force attractive, respectivement répulsive, appliquée entre une paire d'agents, les attirant l'un vers l'autre, ou les repoussant. A chaque pas de temps, un agent "document" remet à jour l'intensité et la direction de cette force en fonction de la dynamique du système (positionnement des documents voisins). Des expériences ont montré que moins le rayon définissant l'espace de voisinage est important, plus les temps de calcul augmentent, la vitesse de convergence diminue, ainsi que l'efficacité du comportement global [HCB13]. Nous avons adapté les travaux de Hong [HCB13] aux agents "document". Notre approche consiste en la sélection d'un agent de manière aléatoire

dans l'environnement de l'agent "document". Ce voisinage est défini de manière temporaire selon un ou plusieurs pas de temps et ne repose pas sur une proximité spatiale. Chaque agent, connaissant la position de son voisin, se déplace de manière interactive en fonction des forces calculées sur cet intervalle de temps et sur cet espace de voisinage.

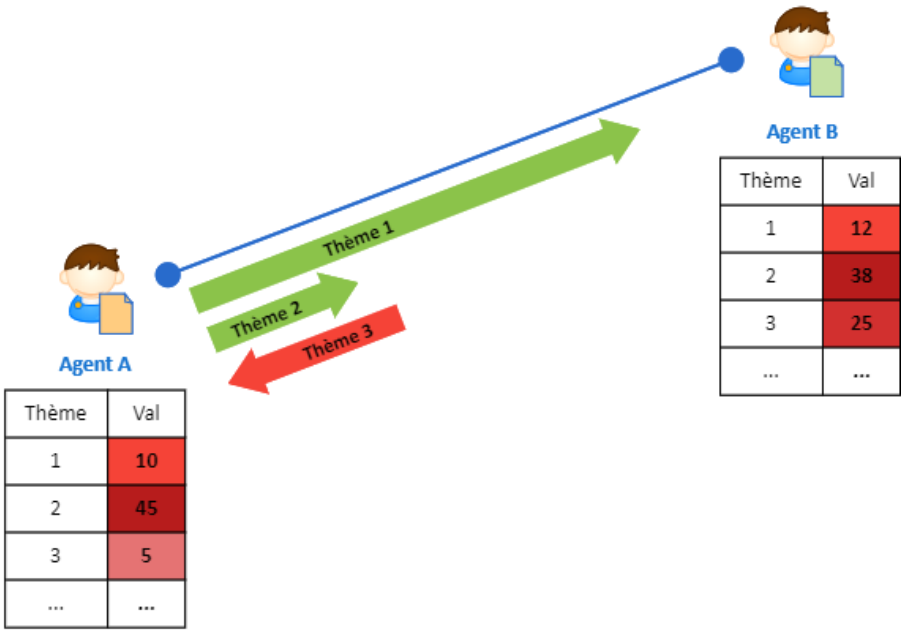


FIGURE *3.8* – ***Chaque agent "document" a un vecteur caractéristique qui décrit les liens du document avec chaque thème. Ces liens servent à calculer les forces d'attraction et répulsion de l'agent "document" A. La somme des forces résultantes s'ajoute à celles calculées aux itérations précédentes.***

Similarité entre paires d'agents "document". La similarité S entre deux agents "document" est calculée par la formule :

$$S^n_{d_a,d_b} = 1 - \frac{|C^n_{d_a} - C^n_{d_b}|}{V^n_{max} - V^n_{min}} \tag{3.2}$$

Où $S^n_{d_a,d_b}$ représente la similarité entre la paire de documents d_a et d_b pour le thème n. $C^n_{d_a}$ correspond à la valeur de *tf-idf* du document d_a (cf. équation 3.1). V^n_{max}, respectivement V^n_{min}, représente la valeur maximale, respectivement la valeur minimale de la composante C de l'ensemble des documents pour le thème n. Un résultat proche de 1, respectivement proche de 0, indique une similarité, respectivement dissimilarité, pour la paire de documents.

Distance entre paires d'agents "document". La distance d entre deux agents "document" est calculée à l'aide d'une distance euclidienne :

$$d_{d_a,d_b} = \sqrt{(x_{d_a} - x_{d_b})^2 + (y_{d_a} - y_{d_b})^2 + (z_{d_a} - z_{d_b})^2} \tag{3.3}$$

Coefficient de force entre paires d'agents "document". Le coefficient de force V pour l'agent "document" d_a en fonction de l'agent d_b est calculée par la formule :

$$V_{d_a,d_b} = \frac{1}{|n|}(\sum_{i=1}^{n} \frac{S^i_{d_a,d_b} * d_{d_a,d_b}}{SS} - t) \tag{3.4}$$

Où V_{d_a,d_b} représente le coefficient de force appliqué à l'agent "document" d_a. SS est un facteur de taille de la simulation et est égale à $0.5*|D|$, D correspondant à l'ensemble des agents. Elle permet de définir les limites de déplacement des agents en fonction de leurs nombres. t correspond au seuil d'attraction/répulsion pour lequel une valeur proche de 0, respectivement proche de 1, favorise l'attraction, respectivement la répulsion, entre les agents. Le résultat compris entre -1 et 1 indique le sens d'orientation de la force appliquée à l'agent "document" d_a.

Force entre paires d'agents "document". La force F appliquée à l'agent "document" d_a est calculée par le système :

$$F_{d_a,d_b} = \begin{cases} x &= (x_{d_b} - x_{d_a}) * V_{d_a,d_b} \\ y &= (y_{d_b} - y_{d_a}) * V_{d_a,d_b} \\ z &= (z_{d_b} - z_{d_a}) * V_{d_a,d_b} \end{cases} \tag{3.5}$$

Cette force est ajoutée à la position de l'agent "document". Les valeurs SS et t ont une valeur arbitraire de 0.5. Pour éviter une congestion liée à l'accès à la liste des agents disponibles, un agent sélectionne un voisin pour lequel sont calculées des forces durant plusieurs itérations. La figure 3.9 décrit les forces qui attirent et repoussent deux agents "document". Cette force est appliquée à l'agent qui définit sa similarité par rapport à un agent de son voisinage. Sa direction suit la droite qui passe par les deux agents. Plus la distance est grande, moins les documents sont similaires.

3.3.2.4 Requêtes documents dans le système multi-agents

Il était nécessaire dans cette étude de développer une interaction simple et intuitive avec les documents. Nous utilisons le système d'attraction/répulsion à cet effet. Les utilisateurs, par le simple clic sur un document/agent, peuvent créer un agent appelé requête. Cet agent n'est alors plus influencé par les forces d'attraction/répulsion. L'utilisateur peut le déplacer dans l'espace, réorganisant par conséquent l'affichage des résultats (cf. Figure 3.10), les forces d'attraction/répulsion étant toujours opérationnelles pour les autres agents. Cet agent-requête statique produit en effet des forces d'attraction/répulsion vers les autres agents qui se réorganisent (cf. Figure 3.11). Ces requêtes, positives ou négatives sont similaires au retour de pertinence utilisé dans les systèmes CBIR traditionnels, produisant un mécanisme simple d'intéraction. Les mots-clés

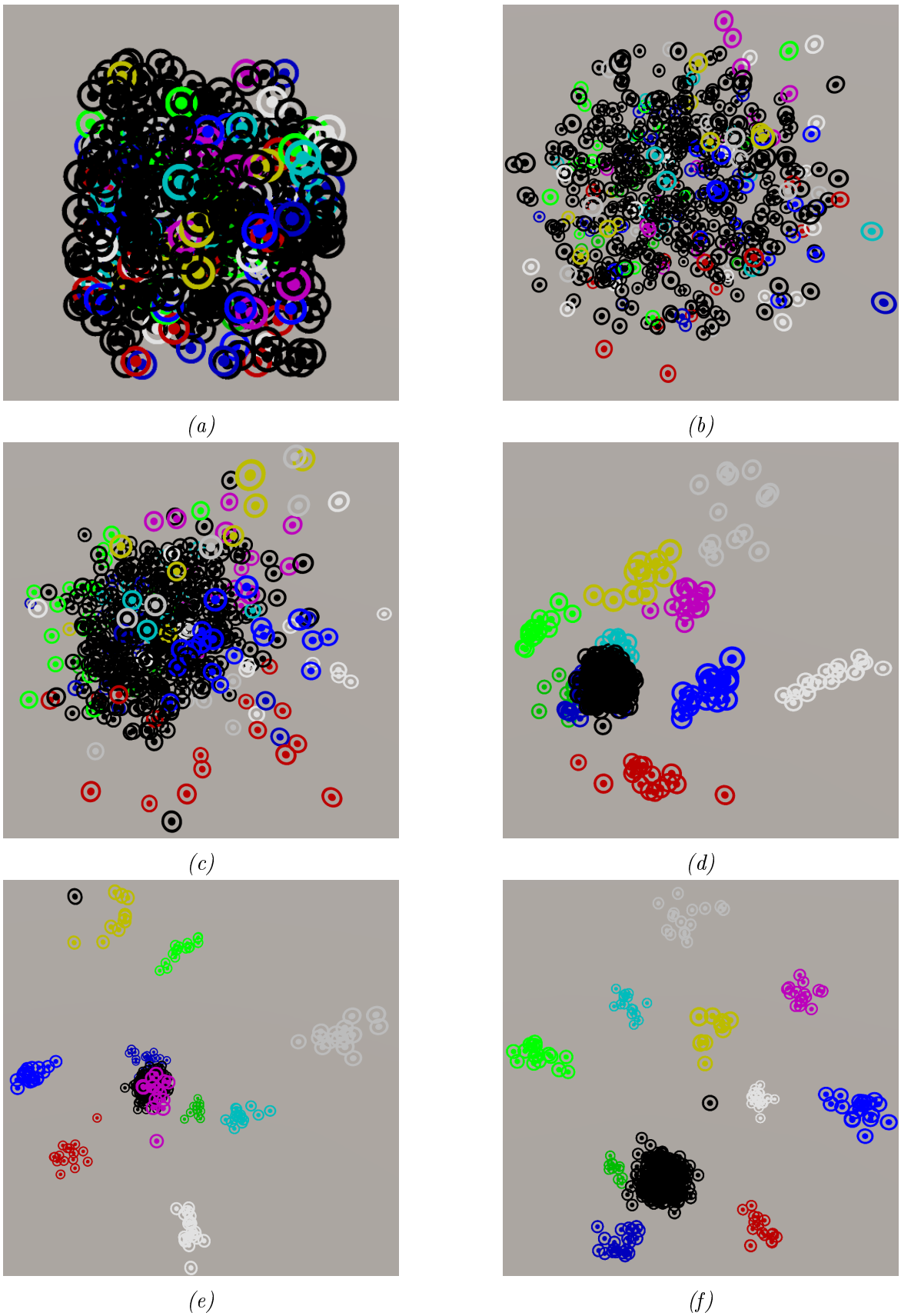


FIGURE *3.9* – ***Représentation sous Unity de l'évolution des agents "document" dans un environnement 3D en fonction des forces d'attraction/répulsion qui les animent. La figure 3.11a représente les agents dans leur état initial. Grâce aux forces de similarité/dissimilarité entre agents, ceux-ci commencent par s'éloigner les uns des autres (cf. figures 3.11b et 3.9c) puis se regroupent entre agents "document" similaires, et commencent à former des clusters (cf. figure 3.9d), pour enfin tomber dans des états stables (cf. figures 3.9e et 3.9f).***

présents dans le document sélectionné peuvent aussi servir à orienter les recherches sur le web ou le corpus et guider ainsi les agents de recherche.

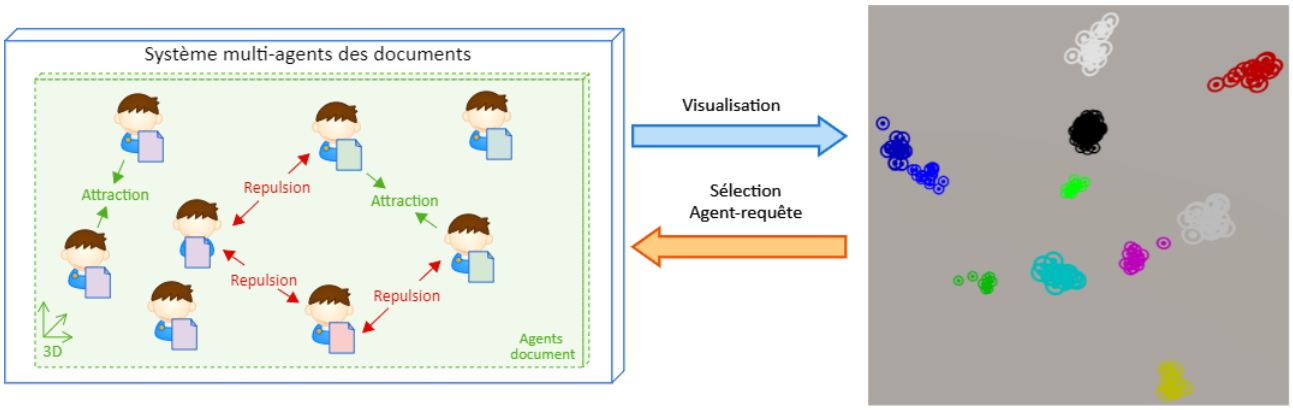
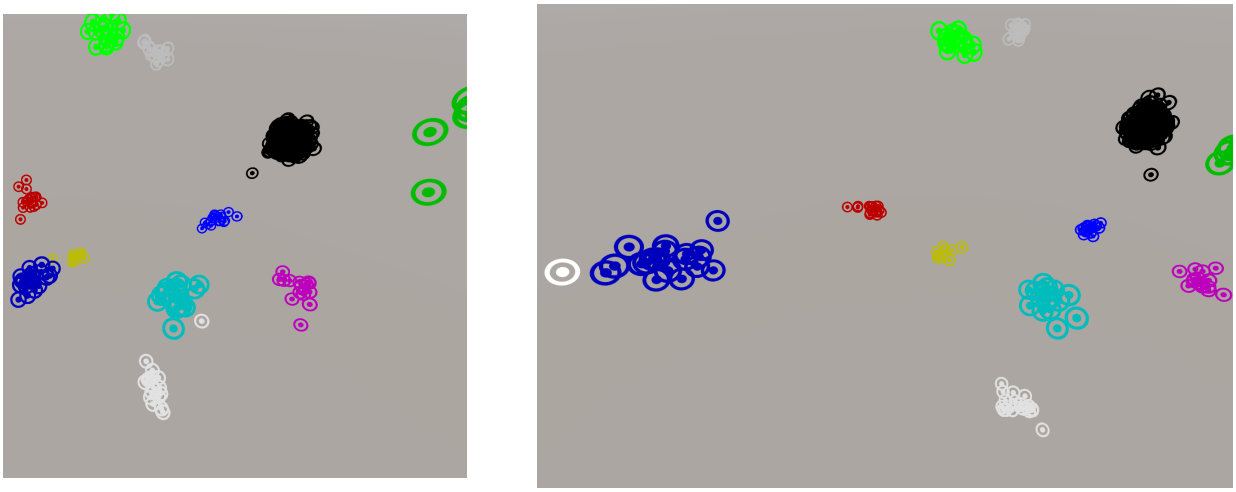


FIGURE *3.10* – ***L'utilisateur visualise les documents du système multi-agents dans un espace en 3D. Chaque document est représenté par un point dans l'espace. L'utilisateur peut sélectionner un agent pour le transformer en agent-requête. Les autres agents de l'espace vont s'organiser autour de cet agent que l'utilisateur aura sélectionné en utilisant les forces d'attraction/répulsion calculées entre paires d'agents à partir des N composantes des vecteurs caractéristiques.***



(a) *(b)*

FIGURE *3.11* – ***Représentation sous Unity d'un agent-requête (couleur blanc). La figure 3.11a représente les agents dans leur état stable. Un agent de couleur bleu est transformé en agent-requête : il génère des forces d'attraction/répulsion qui attirent/repoussent les autres agents (cf. figure 3.11b).***

3.3.3 Système multi-agents de recherche

Dans le système multi-agents associé à la recherche de nouveaux documents évoluent des agents de recherche ne nécessitant pas de représentation dans l'espace d'intéraction. Ils utilisent

un algorithme génétique, s'appuyant sur un critère de pertinence, pour générer de nouveaux agents de recherche.

3.3.3.1 Agent de recherche

Un agent de recherche est un processus représentant une requête composée de plusieurs mots que nous appelons mots-clés, lancée dans un moteur de recherche (cf. Figure 3.12). Un agent de recherche n'a pas de représentation graphique et n'est pas influencé par les forces d'attraction/répulsion des agents "document". Pour chaque thème, plusieurs agents de recherche effectuent des requêtes composées de un ou plusieurs mots. L'ordre des mots d'une requête ayant une influence sur les résultats de la recherche, toutes les combinaisons de mots sont utilisées par l'agent. L'ensemble des résultats est collecté par cet agent qui effectue alors les requêtes nécessaires à la récupération des pages Web. Pour chacune d'elles, l'agent récupère, dans le contenu de la page, la position de tous les mots-clés qui ont servi à la recherche sur le moteur de recherche.

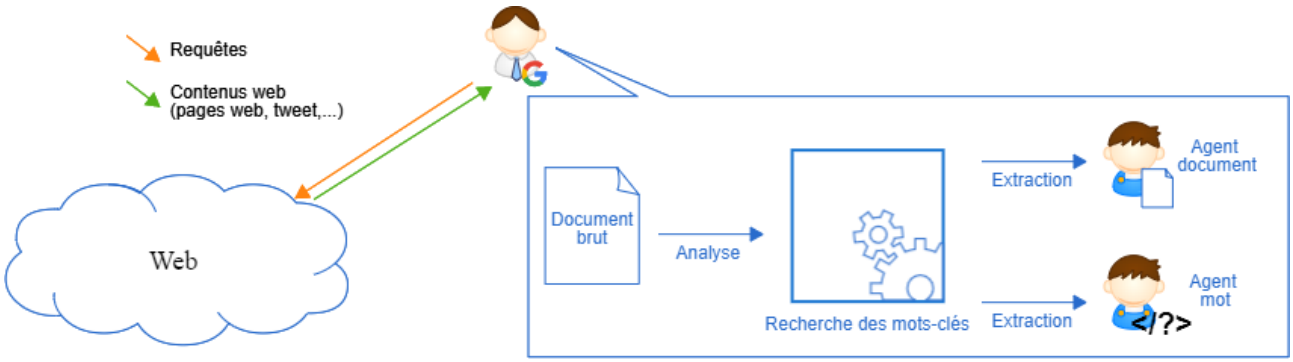


FIGURE *3.12* – ***Un agent de recherche effectue des requêtes dans un moteur de recherche avec des mots-clés d'un thème (e.g.*** signal faible ***potentiel). Pour chaque document trouvé, il extrait le segment de document le plus pertinent contenant les mots-clés de la requête et l'ajoute aux documents déjà présents dans le corpus d'étude afin d'enrichir les données.***

Extraction de nouveaux documents

A partir de la position des mots-clés dans le document récupéré, l'agent recherche la section la plus pertinente contenant l'ensemble des mots-clés (cf. figure 3.13). Cette section de texte est alors transformée en agent "document".

Le système s'enrichit donc de nouveaux documents par ces recherches d'information complémentaire. Ils sont intégrés aux corpus d'étude et transformés en agents "document". Le système mis en place permet donc d'étudier la diffusion / d'investiger des mots du thème *signal faible* potentiel pour fournir aux décideurs un retour d'information sur sa pertinence et son évolution.

Extraction de mots voisins / création de nouveaux agents "mot"

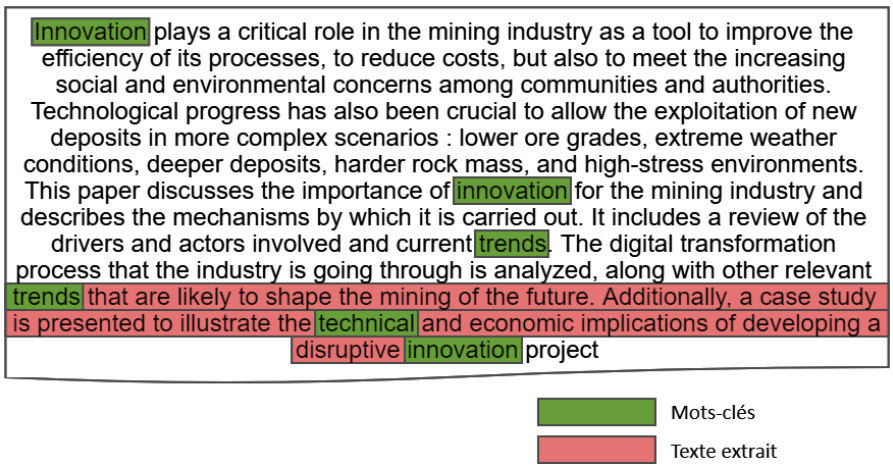


FIGURE *3.13* – ***L'agent de recherche identifie la position des mots-clés (en vert) dans le document. Il extrait du document le segment de texte le plus pertinent contenant l'ensemble des mots-clés. Dans cet exemple, la requête était composée des mots-clés "innovation", "trends" et "technical". La section de couleur rouge correspond à la section récupérée pour former un nouvel agent "document".***

Pour chaque nouveau document extrait par un agent de recherche, ce dernier enregistre les mots situés au voisinage des mots-clés (mots utilisés comme requête sur le moteur de recherche). Sur la figure 3.14, nous illustrons le cas où les requêtes sont composées seulement de 1 mot-clé. Pour chaque combinaison du mot-clé avec un ou plusieurs mots du segment extrait, une entrée dans la liste est créée et sa valeur d'incrément augmentée suivant les résultats donnés par les recherches successives effectuées par chaque agent. L'ensemble des mots associés au mot-clé représente son voisinage. Pour le système multi-agents associé aux mots, le calcul des forces d'attraction/répulsion s'appuie sur ces informations. Si le mot n'existe pas parmi les agents "mot", celui-ci est créé.

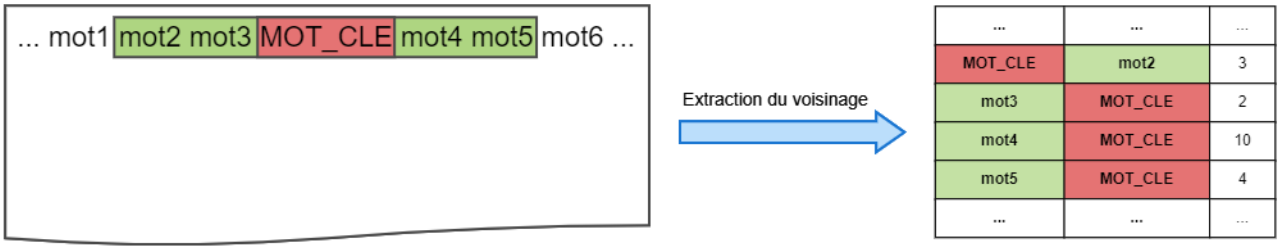


FIGURE *3.14* – ***Pour chaque combinaison du mot-clé avec un ou plusieurs mots du segment extrait, une entrée dans la liste est créée et sa valeur d'incrément augmentée suivant les résultats donnés par les recherches successives effectuées par chaque agent.***

Recherche de mots inconnus dans *Word2Vec*

L'agent de recherche vérifie que les mots situés au voisinage des mots-clés existent dans les vecteurs de mots connus de *Word2Vec*. Dans le cas où le mot ne serait pas présent, l'agent

de recherche crée un nouvel agent de recherche contenant uniquement ce mot inconnu afin de récupérer des documents permettant sa mise en contexte dans *Word2Vec*. Ces documents, servant uniquement à entrainer *Word2Vec*, ne sont pas ajoutés au système multi-agents associé aux documents.

Sélection génétique des futurs agents de recherche

Dans notre processus de recherche de nouveaux documents pour l'enrichissement des données via le système multi-agents, nous mettons en œuvre un algorithme génétique. Avec cet algorithme, nos recherches ne sont pas effectuées avec des mots sélectionnés aléatoirement dans un thème, mais ils sont choisis en s'appuyant sur des précédentes recherches. Dans notre implémentation, un agent de recherche représente pour l'algorithme génétique un individu, un mot-clé associé à une recherche, définit un gène et une recherche composée de plusieurs mots est un chromosome. Dans notre phase de sélection, nous utilisons un critère de pertinence à optimiser (précisé plus loin dans cette section). On génère ensuite de façon itérative des populations d'individus sur lesquelles on applique des processus de mutation. Les agents de recherche récupèrent le groupe de mots-clés (chromosome) provenant d'une précédente recherche effectuée par un agent de recherche (individu) et pour lesquels le critère de pertinence donne la meilleure valeur parmi les recherches effectuées. Ils appliquent des mutations remplaçant un ou plusieurs mots (gènes) par des mots appartenant au même thème. L'ordre des mots ayant une importance pour les moteurs de recherche sur le Web, l'agent de recherche effectue toutes les combinaisons possibles afin de récupérer le maximum de résultats pertinents.

Lorsqu'une recherche donne un grand nombre de résultats, il s'avère pertinent d'effectuer des recherches proches en utilisant les mêmes mots-clés puis de changer un des mots par un autre présent dans le thème. Ainsi nous concentrons les quelques recherches qui suivent par des requêtes proches.

Pour juger de la pertinence des résultats et décider de continuer ou non les recherches avec certains de ces mots-clés, nous définissons un critère prenant en compte les documents ajoutés au système multi-agents par l'agent de recherche car ils contiennent tous les mots-clés de la recherche. L'agent de recherche ayant à effectuer cette recherche pour un thème, nous calculons la somme des *tf-idf* des mots présents dans les documents qui appartiennent à ce thème (cf. formule 3.6) afin de déterminer la pertinence du chromosome.

$$Pertinence(d_a, c_n) = \sum_{i=1}^{N} tf\text{-}idf_i \tag{3.6}$$

Dans la formule 3.6, $Pertinence(d_a, c_n)$ représente la pertinence du document d_a pour le thème c_n, $tf\text{-}idf_i$ la valeur de pondération du mot t_i dans le corpus D, i parcourant les mots présents dans le document d_a. Puis cette somme est calculée pour tous les documents ajoutés au système multi-agents par l'agent de recherche.

Cette valeur de pertinence permet d'évaluer le chromosome par rapport aux documents qu'il a rapportés. Pour éviter qu'un même chromosome soit toujours choisi (aucune autre combinaison ne donne par exemple une meilleure valeur du critère) nous utilisons un système "d'oubli". Après un certain nombre d'itérations de recherche sur la base de ce chromosome (dans les expérimentations, nous avons pris la valeur 10), le système supprime celui-ci. Les futurs recherches utiliseront alors des mots aléatoires choisis parmi ceux présents dans le thème. Si le nouveau chromosome construit donne une valeur de pertinence plus forte, le chromosome précédent est définitivement supprimé, remplacé par celui-ci et le compteur d'itération est remis à zéro.

3.3.4 Système multi-agents associé aux mots

Le système multi-agents associé aux mots s'exerce dans un espace en 3D dans lequel chaque mot-clé, récupéré par les agents de recherche, évolue sous forme d'agent "mot". Ceux-ci sont soumis à des forces d'attraction/répulsion prenant en compte la fréquence d'apparition normalisée des mots présents dans leurs voisinages (cf. section 3.3.3.1). Ce système multi-agents auto-organisé permet à l'utilisateur d'interagir avec les agents en forçant leur position dans l'espace.

3.3.4.1 Agent "mot"

Dans ce système multi-agents, chaque mot référencé dans un thème est représenté par un agent. Les forces d'attraction/répulsion associées à l'agent "mot" sont définies à partir du rapport entre le nombre d'apparition de ce mot au voisinage d'un autre mot et la valeur la plus grande parmi les paires d'agents "mot" présents dans le système (cf. équation 3.7).

Similarité entre paires d'agents "Mot". La similarité S entre deux agents "mot" est calculée par la formule :

$$S_{m_a,m_b} = \frac{Manhattan(m_a, m_b)}{maxManhattan} \tag{3.7}$$

Dans l'équation 3.7, S_{m_a,m_b} représente le rapport de distance pour la paire de mot (m_a, m_b), $Manhattan$ est la distance de Manhattan, $maxManhattan$ est la valeur maximale de cette distance pour toutes les paires de mots.

Distance entre paires d'agents "mot". La distance d entre deux agents "mot" est calculée à l'aide d'une distance euclidienne :

$$d_{m_a,m_b} = \sqrt{(x_{m_a} - x_{m_b})^2 + (y_{m_a} - y_{m_b})^2 + (z_{m_a} - z_{m_b})^2} \tag{3.8}$$

Coefficient de force entre paires d'agents "mot". Le coefficient de force V pour l'agent "mot" m_a en fonction de l'agent m_b est calculée par la formule :

$$V_{m_a,m_b} = \frac{S_{m_a,m_b} * d_{m_a,m_b}}{SS} - t \tag{3.9}$$

Où V_{m_a,m_b} représente le coefficient de force appliqué à l'agent "document" m_a. SS est un facteur de taille de la simulation (similaire à celle pour le calcul des agents "document"). t correspond au seuil d'attraction/répulsion.

Force entre paires d'agents "document". La force F appliquée à l'agent "mot" m_a est calculée par le système :

$$F_{m_a,m_b} = \begin{cases} x &= (x_{m_b} - x_{m_a}) * V_{m_a,m_b} \\ y &= (y_{m_b} - y_{m_a}) * V_{m_a,m_b} \\ z &= (z_{m_b} - z_{m_a}) * V_{m_a,m_b} \end{cases} \tag{3.10}$$

Cette force est ajoutée à la position de l'agent "mot".

3.3.5 Analyse de l'évolution des thèmes

Pour étudier l'évolution des mots d'un thème au cours de l'investigation, un suivi historique est nécessaire afin de (1) sauvegarder l'état du système multi-agents associé aux mots et (2) effectuer la recherche de documents avec une liste pertinente et limitée de mots-clés. Ce système étudie l'évolution des thèmes en conservant un nombre fini de mots associés et fournit des indicateurs de la dynamique observée sur les thèmes *signaux faibles* potentiels.

Etudier l'évolution de thèmes nécessite de relancer régulièrement de nouvelles recherches. Le nombre de mots sauvegardés est limité, notamment pour permettre une recherche réaliste (en termes de temps) utilisant toutes les combinaisons possibles des mots de la liste. Selon la nature du problème, on se limitera à des combinaisons de mots de taille 1 à 3.

Pour l'ensemble des recherches effectuées, le système sauvegarde la date de fin de traitement ainsi que les données permettant de recalculer la valeur *tf-idf* de chaque mot. Les traces successives du déroulement de l'investigation conduit ainsi à l'obtention de graphiques et indicateurs que nous présenterons dans le chapitre 4. Ils permettent une aide à la décision sur le fait que le (ou les) thème porteur d'un *signal faible* potentiel révélé par l'algorithme 2 du chapitre 2 est effectivement (ou non) porteur d'un *signal faible*.

3.4 Présentation de l'architecture du logiciel WILD

La plateforme d'investigation numérique développée dans cette étude est nommée WILD ("Weak sIgnaL Discovery"). Elle met en œuvre une solution logicielle sous forme d'un ensemble de composants et services travaillant en parallèle et utilisant un système de verrous afin de garantir la cohérence des ressources partagées (cf. Figure 3.15). Ces ressources sont verrouillées dans le même ordre pour garantir aucune situation de blocage total (*deadlock*). L'utilisation de multiples unités d'exécution (*threads*) fournit une solution à l'exécution simultanée de plusieurs services, des systèmes d'analyse et de suivi ainsi que des composants système multi-agents travaillant simultanément sur des bases différentes sur une même machine.

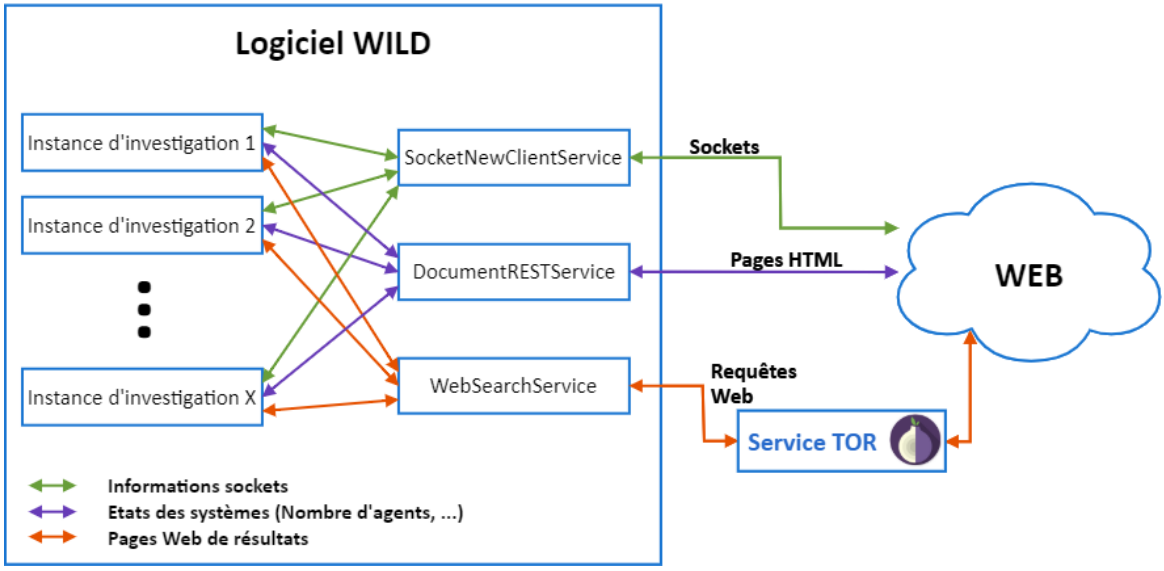


FIGURE *3.15* – ***Vue globale du logiciel WILD et des différents services connectés servant d'interface pour l'accès au Web.***

L'ensemble des traitements présentés a été mis en œuvre sur un ordinateur portable :
— Core i7 (4 coeurs / 8 threads)
— 16 Go de RAM
— Coeur graphique Intel HD 4600
— Windows 10
Et sur lequel sont lancés simultanément :
— Le logiciel WILD
— Le service TOR
— Un client Unity connecté par socket au logiciel
— Un navigateur internet affichant l'interface Web d'administration

Dans les sections suivantes, nous détaillons l'ensemble des composants et services ainsi que l'ordre des traitements et les flux d'informations. Le logiciel est composé de 10 composants / services différents (cf. Figure 3.16). Nous détaillons dans les sections suivantes, leur fonctionnement ainsi que les entités avec lesquelles ils interagissent.

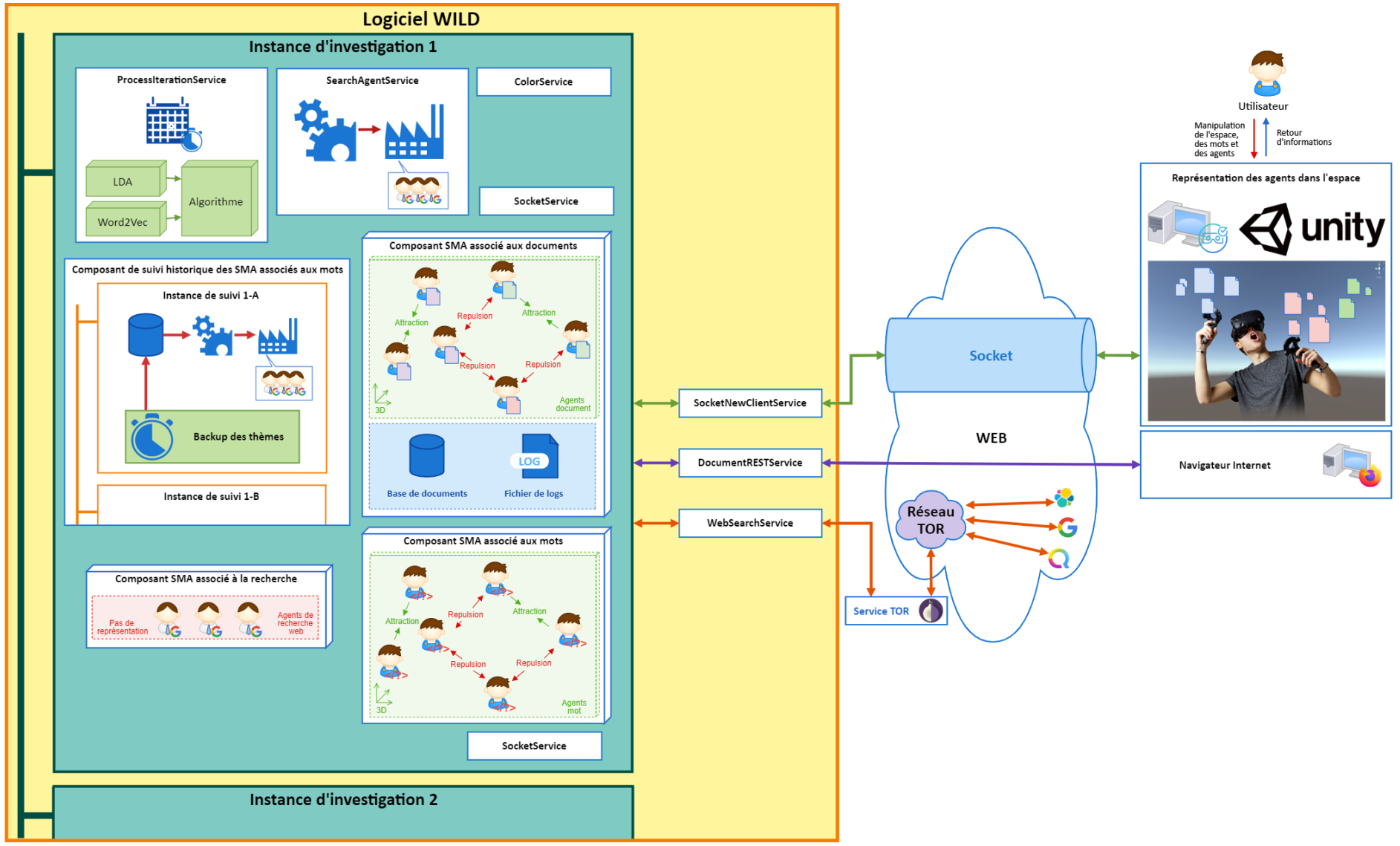


FIGURE *3.16* – ***Vue détaillée de toutes les instances d'investigation, composants et services du logiciel WILD. Sont présentées également, les interfaces d'accès et de recherche : Interfaces Unity et Web et service d'accès aux moteurs de recherche.***

3.4.1 Composants et services du logiciel WILD

3.4.1.1 Instance d'investigation

Plusieurs instances de la chaine complète d'investigation peuvent être lancées en parallèle. Chaque instance forme un processus global exécutant l'ensemble de la chaine de traitement présentée dans les chapitres 2 et 3. Cette instance contient plusieurs composants/SMA/services intéragissant. Les principaux sont le composant SMA (Système Multi-Agents) associé aux documents, le composant SMA associé aux mots et le composant de suivi historique de modèle de mots.

Ces instances partagent un ensemble de services communs s'interfaçant avec d'autres outils (TOR, Client de visualisation sous Unity).

Les instances travaillent en parallèle et disposent chacune de bases de données/documents/-mots/logs spécifiques.

3.4.1.2 Service d'interface Web "DocumentRESTService"

Les utilisateurs tels que les lanceurs d'alerte ou des décideurs doivent pouvoir interagir sur l'instance pour la gestion des paramètres, la mise à jour du corpus de documents ou encore lancer de nouveaux traitements.

Ce service fournit une interface Web permettant à l'utilisateur de gèrer les fonctions globales du logiciel WILD. Elle présente un certain nombre de métriques. Ce service peut être lancé localement ou à distance. Il simplifie l'intervention sur les fichiers de configuration propre au logiciel ou à l'instance.

Pour chaque instance, l'utilisateur peut visualiser un certain nombre de statistiques et exécuter des actions sur celles-ci. Le lanceur d'alerte peut déposer de nouveaux documents depuis l'interface et les utilisateurs réaliser un ensemble de paramétrage tels que :

— la modification du nombre d'agents de recherche dans l'instance ;
— le changement de la plannification du processus itératif *LDA*/*Word2Vec* ;
— l'ajout/le changement/la suppression de la liste des moteurs de recherche utilisables ;
— la mise en place d'une nouvelle sauvegarde des thèmes dans le composant de suivi historique ;
— la création/suppression/mise en pause de l'instance d'investigation ;
— ...

L'interface présente les tableaux de bord :

— le nombre de documents présents dans le corpus de l'instance ;
— le nombre d'agents de recherche ;

— le suivi des thèmes sauvegardés dans le composant de suivi ;
— le nombre de mots présents ;
— ...

3.4.1.3 Service de recherche Web "WebSearchService"

Ce service exécute les requêtes demandées par les agents de recherche sur les moteurs de recherche (cf. Figure 3.17). Ces requêtes Web passent aux travers du service TOR disponible sur la machine ou sur la passerelle du réseau.

Le réseau TOR permet l'anonymat des requêtes effectuées. Ceci empêche la détection de l'adresse IP par les moteurs de recherche, pouvant conduire à l'impossibilité d'obtenir des résultats de recherche durant une période indéfinie (plusieurs heures/jours). Le service TOR modifie les noeuds de routage (par défaut toutes les 10 minutes) ce qui autorise le lancement régulier de requêtes et pallie aux problèmes du blocages par les moteurs de recherche.

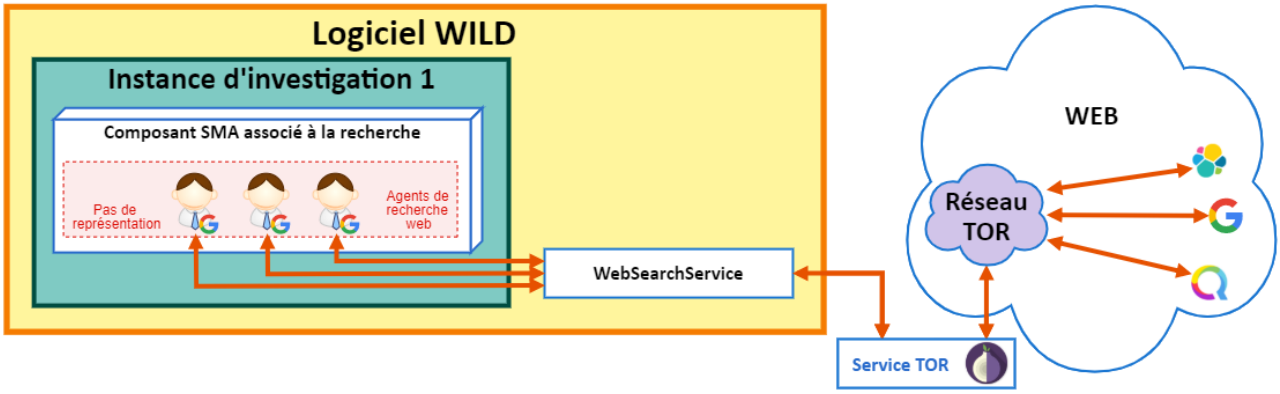


FIGURE *3.17* – ***Présentation des échanges de données entre les agents de recherche, le service "WebSearchService", le service TOR et les moteurs de recherche. Les agents effectuent leurs requêtes auprès du service de recherche Web. Celui-ci lance la requête sur les moteurs de recherche au travers du service TOR.***

Le logiciel WILD propose actuellement deux moteurs de recherche, Qwant et ElasticSearch. Ce dernier permet également de réaliser des expérimentations sur un corpus documentaire dans un environnement controlé. L'intérêt est de maitriser le corpus afin d'évaluer la pertinence des résultats obtenus. D'autres moteurs peuvent être ajoutés. Les plus communs tels que Google ou Bing n'ont pas été investigés dans les travaux.

Les agents de recherche fournissent des groupes de mots au service pour les futures requêtes Web. Une fois la requête effectuée auprès des moteurs, le service retourne à l'agent les liens Web extraits des pages Web résultats. L'agent de recherche récupère les pages Web à partir de ces liens. Celles-ci deviennent alors de nouveaux agents "document" (après extraction du contenu pertinent) ajoutés au composant SMA associé aux documents (cf. Figure 3.18). L'agent effectue

d'autres traitements comme par exemple l'extraction des mots voisins (voir 3.3.3.1) ainsi que la création de nouveaux agents de recherche pour entrainer *Word2Vec* sur des mots non présents dans sa base d'apprentissage initiale (voir 3.3.3.1).

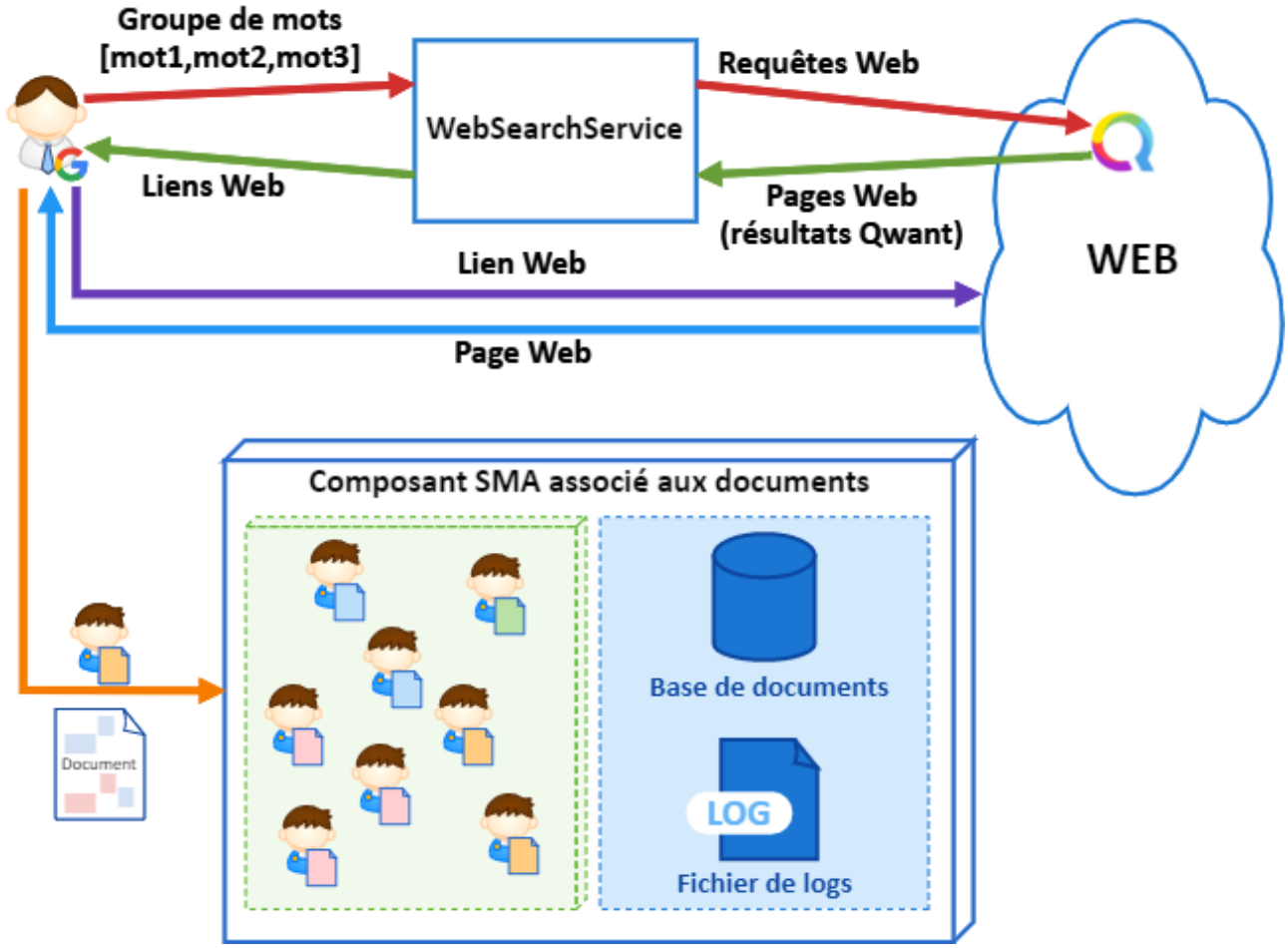


FIGURE *3.18* – ***Un agent de recherche utilise les liens Web fournis par le service de recherche Web, suite à ses requêtes, pour tranformer en agents "document" les pages Web après extraction du contenu pertinent (voir 3.3.3.1)***

3.4.1.4 Service de connexion par socket "SocketNewClientService"

Le logiciel WILD met en œuvre des systèmes multi-agents auto-organisés où les agents "document" ou agents "mot" sont animés par des forces d'attraction/répulsion (voir section 3.3.2 et 3.3.4). Au sein d'un espace 3D, l'utilisateur visualise et manipule ces agents. Cette représentation en temps réel de l'état du système offre une solution d'interaction adaptée à un système en évolution.

L'échange des informations est effectué au travers d'une interface de connexion de type socket. Pour échanger de l'information entre un client de visualisation sous Unity et le logiciel WILD, la solution suit une procédure spécifique. Cette dernière passe par plusieurs phases (cf.

Figure 3.19) :

— le client (application Unity) se connecte au logiciel via un port spécifique. La connexion est alors initialisée dans le logiciel WILD ;

— le client indique le numéro de l'instance pour laquelle il souhaite obtenir les informations à projeter dans l'espace 3D ainsi que son choix dans la visualisation des documents ou des mots.

— le service transfert alors la connexion au service d'échange de données "SocketService" correspondant.

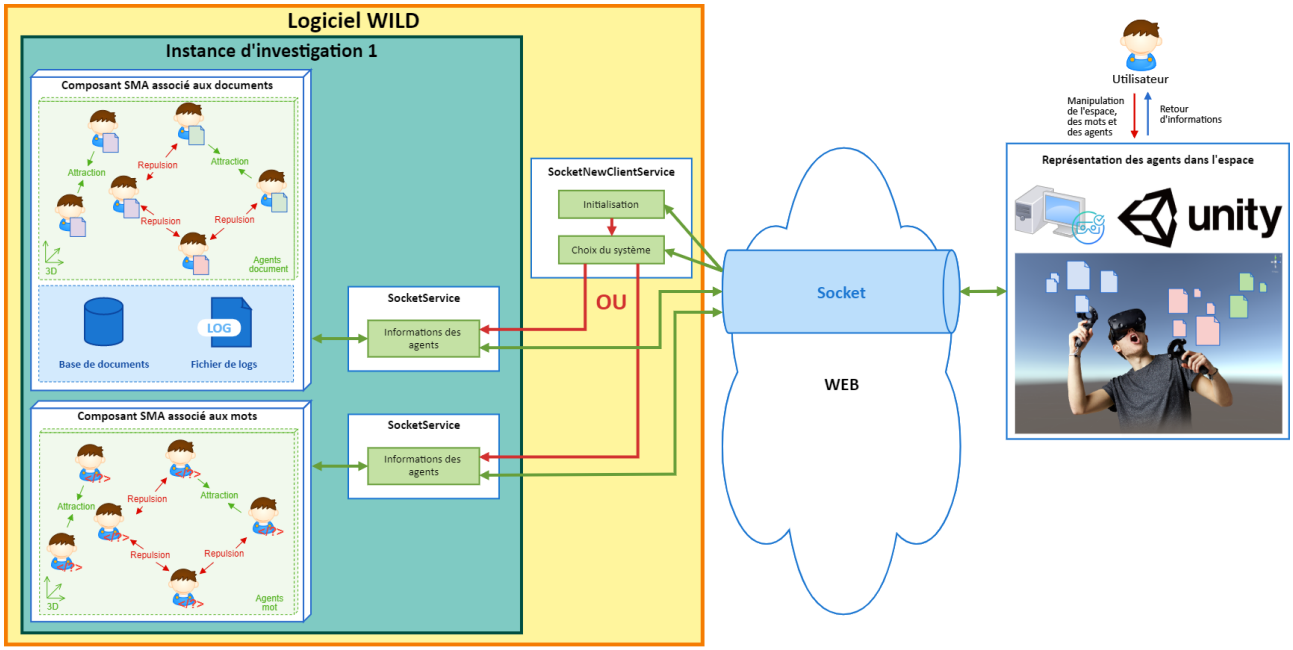


FIGURE *3.19* – ***Présentation des échanges de données via une socket entre un client de visualisation sous Unity et le logiciel WILD. Le service écoute sur un port, attendant la connexion de clients et le choix des informations sur l'instance qu'il souhaite afficher. La connexion est passée au service qui échange ces informations avec le client connecté.***

Les échanges de données s'effectuent aux moyens du format JSON permettant de rapidement sérialiser/désérialiser l'information.

3.4.2 Composants et services du système d'investigation

Chaque instance d'investigation de documents est composé de plusieurs services, d'un composant SMA associé aux documents, d'un deuxième associé aux mots, d'un troisième associé à la recherche et d'un composant de suivi historique des SMA associés aux mots. Cette architecture est répliquée pour chaque nouvelle instance dont les paramètres spécifiques sont modifiables par l'interface Web présentée dans la section 3.4.1.2. Nous présentons ici chacun des composants et services qui compose un tel système d'investigation et leurs intéractions.

3.4.2.1 Composant SMA associé aux documents

Le composant SMA, associé aux documents, est constitué des agents "document" animés par des forces d'attraction/répulsion (voir section 3.3.2). Le système met à jour un nombre d'agents variable et dépendant de la charge CPU disponible de la machine sur laquelle est lancé le logiciel WILD avec la contrainte d'une vitesse d'actualisation fixée arbitrairement à 60 fois par seconde.

Ce composant gère également la base de documents et les logs. Ces derniers retracent le suivi des résultats obtenus et des modifications paramètrant le système multi-agent de documents :

- la liste des nouveaux documents obtenus et les requêtes associées ;
- la mise à jour des thèmes après chaque exécution du processus itératif "ProcessIteration-Service" (voir section 3.4.2.4) ;
- les groupes de mots utilisés comme requêtes sur les moteurs de recherche ;
- les données de suivi historique du modèle de mots (voir section 3.4.2.6) ;
- ...

3.4.2.2 Composant SMA associé aux mots

Le composant SMA associé aux mots se compose d'agents "mot" animés par des forces d'attraction/répulsion basées sur leurs fréquences d'apparitions dans les documents trouvés par les agents de recherche (cf. Figure 3.16). Similaire au composant SMA associé aux documents, un nombre variable de mots est mis à jour selon la charge CPU pour une actualisation de 60 fois par seconde.

Dans ce composant, seul le nombre d'apparition des paires de mots est sauvegardé dans les logs. Lors du démarrage du logiciel, l'ensemble des mots des thèmes sont chargés dans le modèle multi-agents de mots.

3.4.2.3 Composant SMA associé à la recherche

Le composant SMA associé aux mots se compose d'agents de recherche n'intéragissant pas entre eux et créés par le service d'agents de recherche (cf. section 3.4.2.5) et le composant de suivi historique (cf. section 3.4.2.6). Similaire au composant SMA associé aux documents, un nombre variable de mots est mis à jour selon la charge CPU pour une actualisation de 60 fois par seconde.

Ce composant sauvegarde, dans les logs, la liste des recherches déjà effectuée pour chaque lancement du processus itératif (cf. section 3.4.2.4)

3.4.2.4 Service du processus itératif "ProcessIterationService"

Le processus itératif agit notamment comme un orchestrateur planifiant les exécutions de la phase de modélisation thématique multi-niveaux présentée au chapitre 2 (cf. Figure 3.20). La planification utilise le formalisme de périodicité des expressions cron. Ce service accède nécessairement à la base de documents (cf. Figure 3.21). Le service possède également les fonctionnalités de mise à jour du composant SMA associé aux mots avec comme paramètres les résultats obtenus.

3.4.2.5 Service d'agents de recherche "SearchAgentService"

Il est nécessaire d'effectuer un travail d'enrichissement de l'information dans un objectif d'ouverture à un contexte d'information plus large. Ce service sélectionne des mots au sein des thèmes pour produire des agents qui seront en charge des requêtes transmis au service de recherche Web (cf. section 3.4.1.3) pour l'obtention de nouveaux documents pertinents (cf. Figure 3.22). Le service gère également le nombre d'agents de recherche en cours. Il a en charge la suppression des agents ayant fini leurs tâches. Afin d'assurer la production de nouvelles recherches, le service enregistre les groupes de mots déjà utilisés dans les logs pour ne pas relancer de recherche identique.

L'arrêt des recherches s'effectue quand toutes les combinaisons ont été réalisées (la vérification de cette condition est effectuée par ce service). L'ajout de nouveaux mots ou le changement de thèmes (après traitement du service "ProcessIterationService") relance la création de nouvelles combinaisons de groupe de mots et la création de nouveaux agents de recherche.

Pour chaque thème, le service sélectionne des mots selon leur pondération *tf-idf*. Un nouvel agent "document" est créé à partir de ce groupe de mots. Enfin, il est ajouté au composant SMA associé à la recherche (cf. Figure 3.23).

3.4.2.6 Composant de suivi historique des SMA associés aux mots

Les nouveaux mots récupérés lors des ajouts de nouveaux documents par les agents de recherche (cf. section 3.3.3.1) ainsi que par les exécutions planifiées de l'étape de modélisation thématique multi-niveaux mise en œuvre dans le service du processus itératif (cf. section 3.4.2.4) modifient les thèmes et les listes de mots-clés (ainsi que l'ordre éventuel de ceux-ci, puisque leur valeur de *tf-idf* change). Le logiciel WILD fournit un composant de sauvegarde des listes de mots-clés des thèmes avant leurs modifications. L'utilisateur peut ainsi suivre l'évolution de l'investigation et effectuer de nouvelles recherches (cf Figure 3.24).

Le composant sauvegarde les n premiers mots de chaque thème (dans les expérimentations présentées au chapitre 4, nous prenons $n = 20$). Pour chacun de ces thèmes, le composant

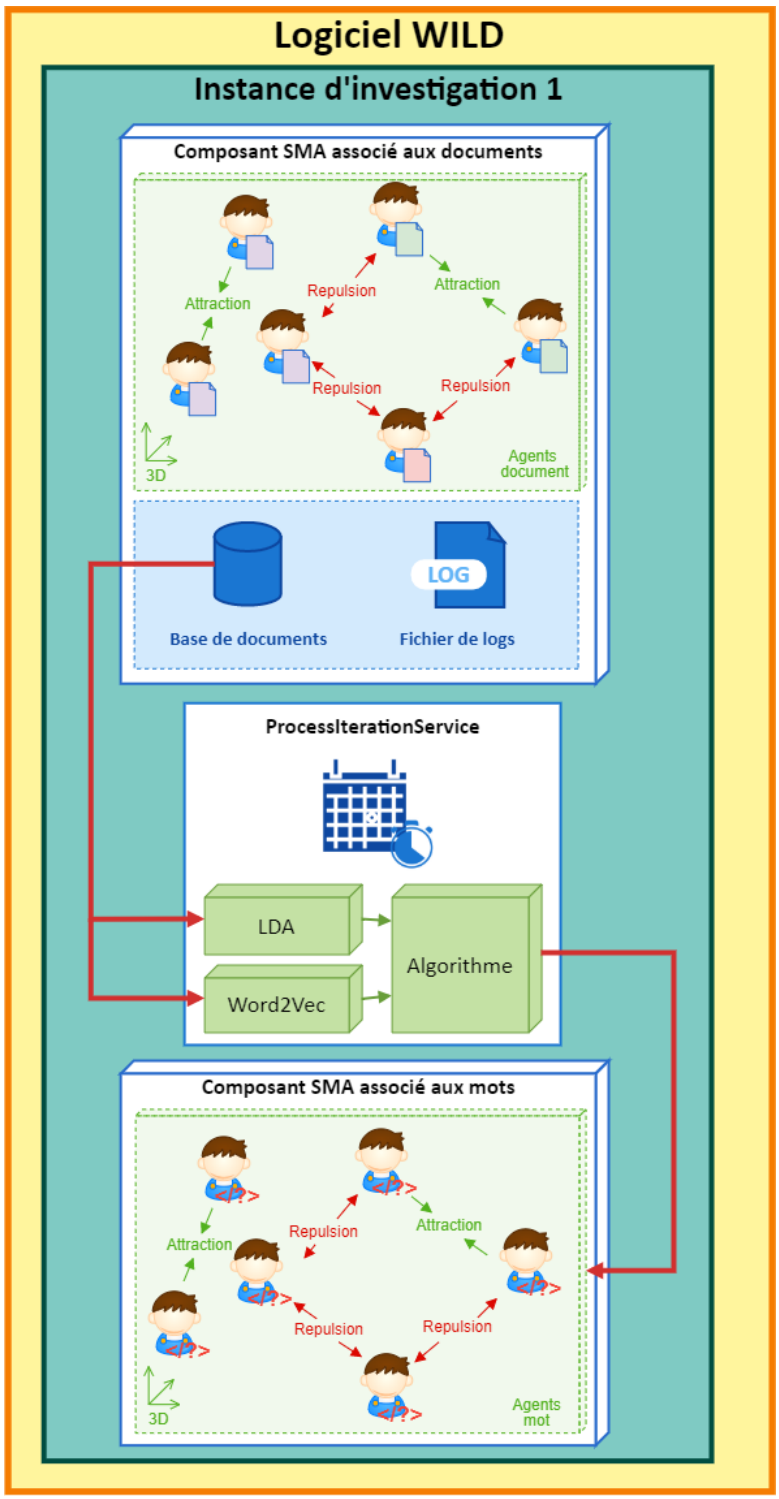


FIGURE *3.20* – ***Le service du processus itératif agit comme un orchestrateur et planifie la phase de modélisation thématique. Les thèmes sont mis à jour et les mots associés à ces thèmes sont de nouveau injectés dans le composant multi-agent associé aux mots.***

lance des recherches à partir de combinaison de groupe de mots de la liste du thème. Une fois l'ensemble des combinaisons exploitées, le composant sauvegarde les nouvelles valeurs *tf-idf* des

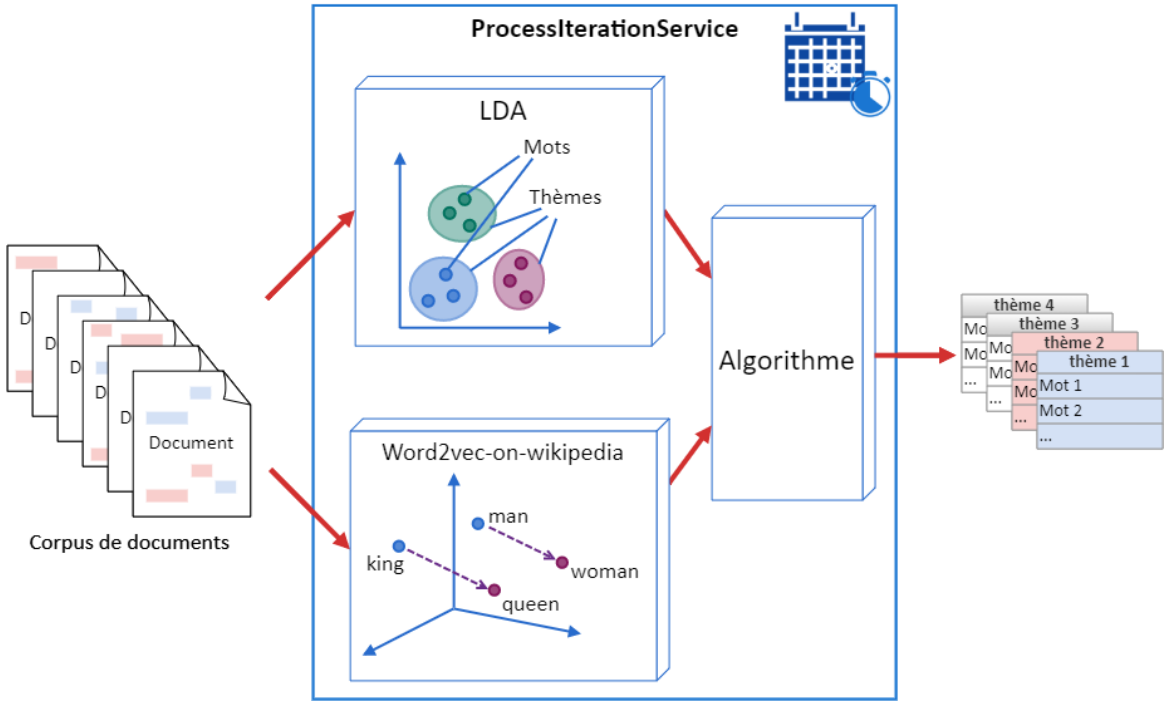


FIGURE *3.21* – ***Le service prend en entrée un ensemble de documents contenant un ou plusieurs*** signaux faibles ***potentiels. Les thèmes sont mis à jour et les listes de mots associés sont de nouveaux injectées dans le composant multi-agent associé aux mots.***

mots (cf. Figure 3.25). Ceci permet de suivre l'évolution des thèmes pour confirmer ou non si le thème est associé à un *signal faible*.

3.4.2.7 Service d'échange de données "SocketService"

Pour chaque composant SMA (associé aux documents ou mots), un service d'échange de données est associé. Ce service prend en charge 2 fonctionnalités :

— l'envoi des positions de chaque agent et de leur vitesse au format JSON à toutes les interfaces de connexion de type socket.

— la récupération des actions envoyées par le client Unity au format JSON. Un flag portant le nom de l'action permet de choisir le traitement approprié.

Ce service gère les agents-requêtes à appliquer au composant SMA en cours de visualisation dans le client Unity.

3.5 Critères et paramétrage de la chaine de traitement

Il est nécessaire d'ajuster un ensemble de paramètres pour chercher à obtenir des résultats cohérents et pertinents. La chaine de traitement décrite sur la figure 3.26 présente la manière

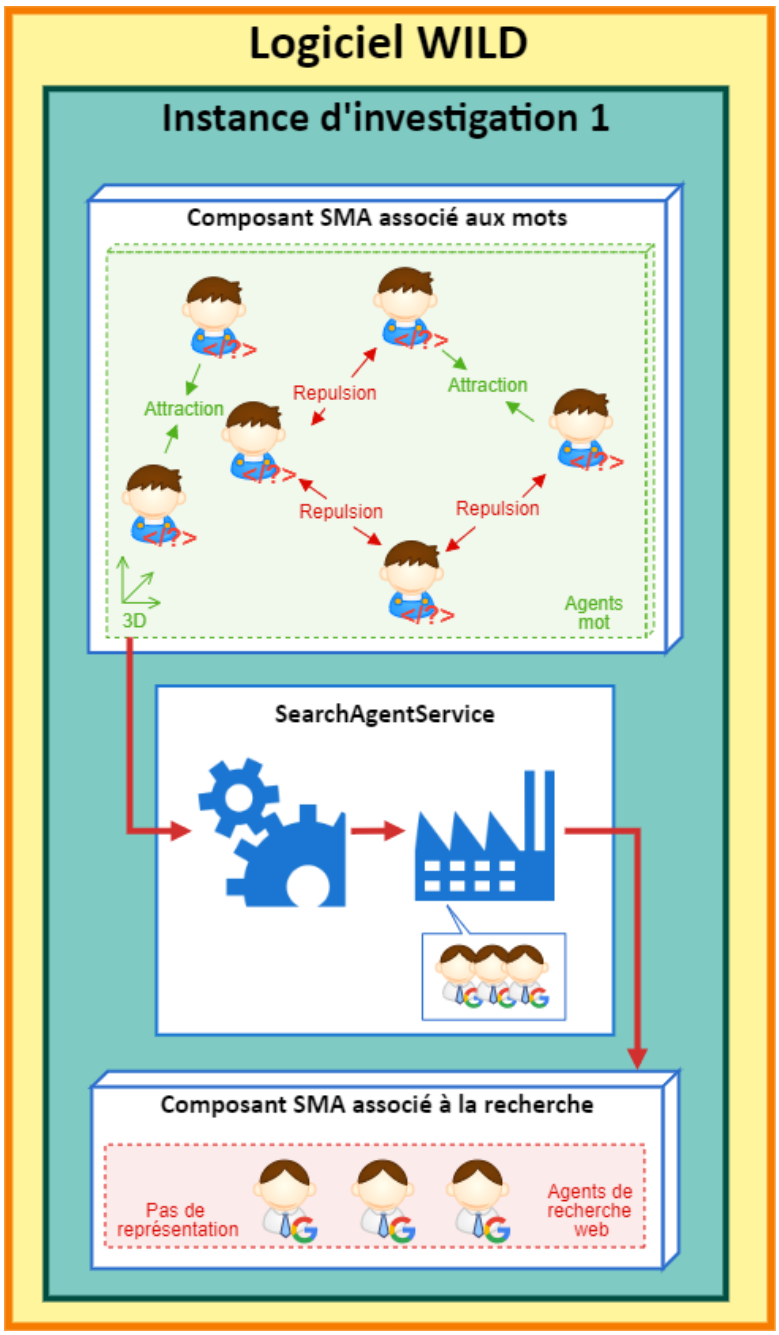


FIGURE *3.22* – ***Le service d'agents de recherche produit, à partir de groupes de mots provenant des thèmes, de nouveaux agents de recherche. Ces agents sont en charge de la récupération des sections de texte pertinentes afin d'augmenter le corpus initial. Ce corpus est ensuite ré-analysé par le service du processus itératif (cf. section 3.4.2.4) qui met en œuvre l'approche de modélisation thématique multi-niveaux (approche conjointe LDA/Word2Vec). Les thèmes et leurs mots-clés associés sont alors mis à jour et de nouvelles recherches sont lancées.***

dont les données circulent et les différents blocs qui la composent :

— ***Agent-based Data Mining*** : Cette section de notre chaine correspond à la phase de modélisation thématique multi-niveaux et à la phase de recherche de nouveaux docu-

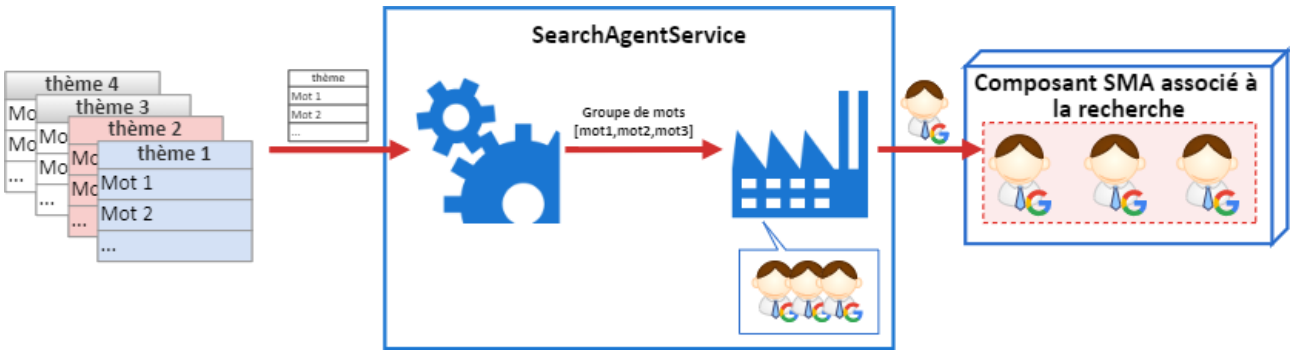


FIGURE *3.23* – ***Le service prend en entrée chaque thème et leur liste de mots associées, et produit des combinaisons de groupe de mots afin de créer de nouveaux agents de recherche. Ces agents transmettent les groupes de mots sous la forme de requête au service de recherche Web (cf. section 3.4.1.3) pour récupérer les résultats de recherche. Les sections de texte résultantes des pages Web apparaissant dans les résultats de recherche définissent alors de nouveaux agents "document".***

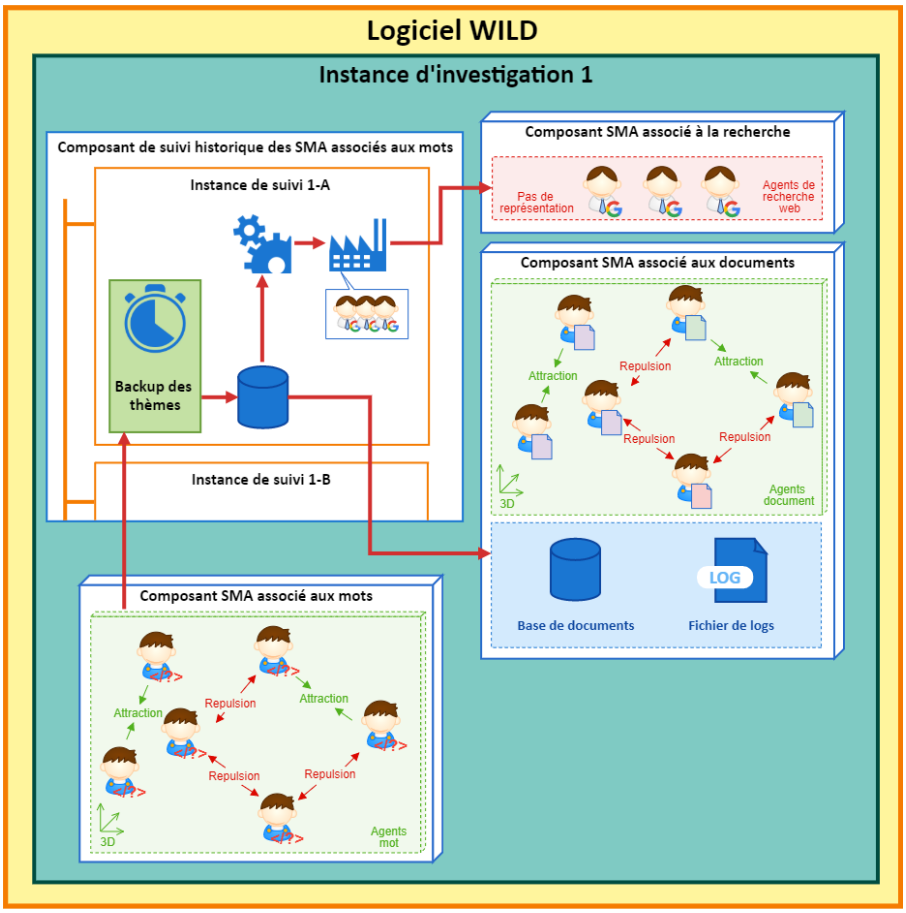


FIGURE *3.24* – ***Le composant effectuant le suivi historique sauvegarde les thèmes et leur liste de mots-clés afin d'étudier leur évolution. Ces mots-clés permettent la construction de nouveaux agents de recherche.***

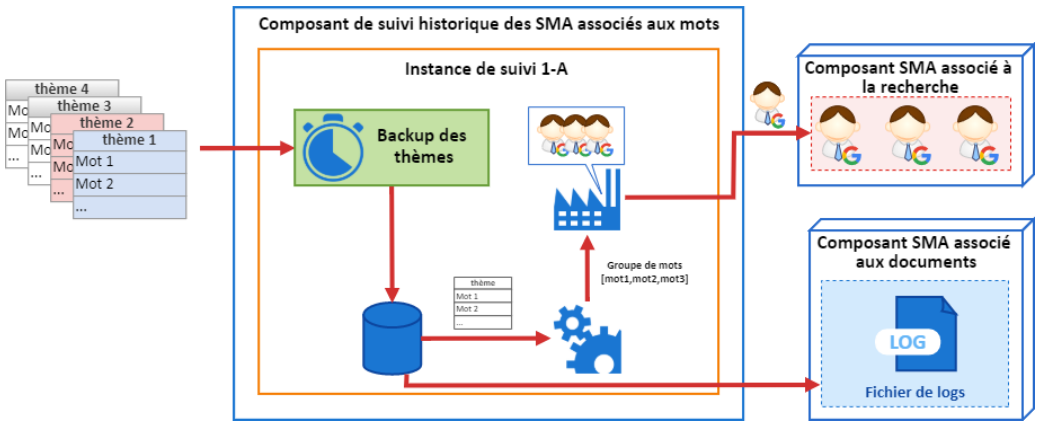


FIGURE *3.25* – ***Le composant effectuant le suivi prend en entrée les thèmes et les listes de mots-clés apportés par le composant SMA associé aux mots. Pour chacun d'entre eux, des agents de recherche sont créés à partir de combinaison de groupes de mots. Le composant enregistre l'évolution des valeurs*** tf-idf ***dans les logs.***

ments au moyen du Web Mining par des agents.
La section *Agent-based Data Mining* décrit des agents travaillant avec différentes sources de données. Ces sources de données sont des moteurs de recherche. Leurs disponibilités sont très variables et la solution qu'est le réseau TOR ne permet pas d'assurer un accès toujours fiable à des résultats de requêtes. Dans notre système, les agents collaborent selon un algorithme génétique, après avoir opéré indépendamment sur des données recueillies, pour partager de la connaissance et planifier les futurs tâches de recherche pour d'autres agents.

— **Agents de gestion de la visualisation** : Les documents sont représentés par des agents évoluant dans une projection 3D animée par des forces d'attraction/répulsion. Cette projection est fournie à l'utilisateur sous la forme d'une interface homme/machine. L'utilisateur effectue des paramètrages dans la visualisation par la composition de requêtes (agent-requête) en déplacant des documents dans l'espace 3D (cf. section 3.3.2.4). Les agents "document", de par leurs vecteurs caractéristiques, s'organisent autour de ces requêtes et proposent des résultats.

Afin d'obtenir des résultats pertinents, l'utilisateur doit déterminer par un réglage fin l'ensemble des paramètres du logiciel WILD. Ces paramètres gèrent plusieurs composants et services. Leurs influences sur les résultats obtenus peuvent varier grandement.

Les paramètres n'ayant pas les mêmes facteurs d'importance, nous avons rassemblé l'ensemble des paramètres que l'utilisateur doit choisir pour les différentes phases de la chaine de traitement sur lesquelles ils interviennent et préciser leur importance sur les résultats obtenus dans le tableau 3.4.

Pour nos expérimentations, les réglages sont réalisés de manière empirique. Les résultats obtenus montrent un choix pertinent de jeux de paramètres. Dans le chapitre suivant, nous

détaillerons plusieurs expérimentations avec plusieurs jeux de paramètres différents afin de montrer leur sensibilité vis-à-vis des résultats.

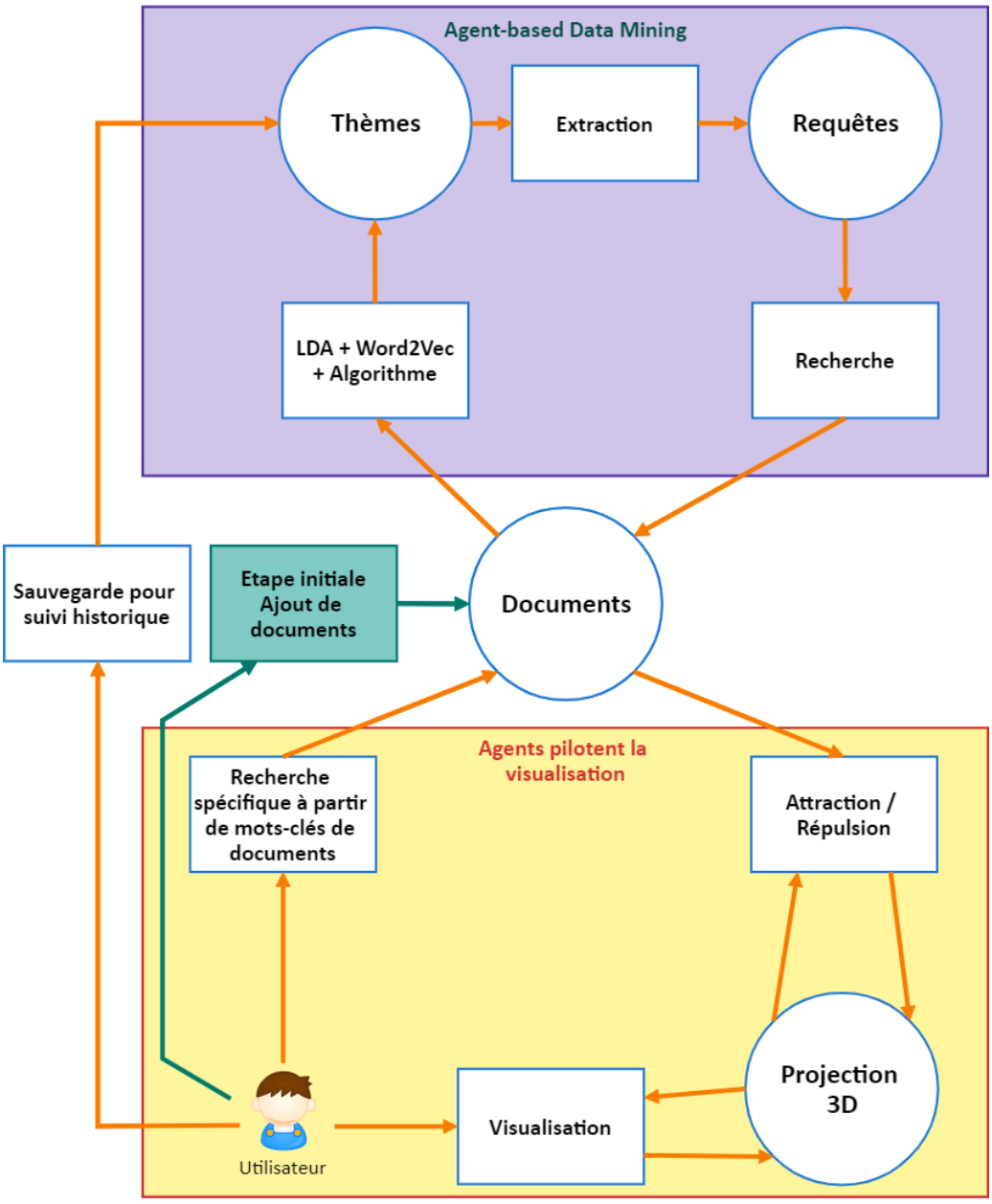


FIGURE *3.26* – ***Présentation fonctionnelle de la chaine de traitement et du sens de transfert des données entre fonctions.***

3.6 Conclusion

Pour la phase d'*agent mining*, combinaison du *data mining* et des systèmes multi-agents, nous proposons une solution pour la recherche d'information et l'interaction des documents avec l'utilisateur. Des agents animés par des forces d'attraction/répulsion se déplacent dans

un espace 3D où l'utilisateur peut formuler des requêtes en figeant et déplacant des agents, réorganisant ainsi l'affichage des autres agents. Des recherches complémentaires, à partir des mots-clés contenus dans les thèmes, enrichissent le corpus initial de nouveaux documents grâce à des requêtes sur des moteurs de recherche. Le suivi des thèmes *signaux faibles* potentiels est réalisé pour étudier leur évolution. Ces résultats sont présentés sous forme visuelle intéractive et tableaux de bord.

Le logiciel WILD présente une solution intégrée de la chaine de traitement décrite dans les précédents chapitres. Cette preuve de concept de plateforme d'investigation fournit un ensemble de services à destination des lanceurs d'alertes et d'utilisateurs dans un objectif de prise de décision. Des algorithmes, adaptés à la recherche de *signaux faibles* et leurs analyses à partir d'approche dynamique et tableaux de bord, sont utilisés dans le logiciel pour traiter, analyser et hiérarchiser l'information hétérogène.

Le logiciel est composé d'un ensemble de services et composants travaillant en parallèle. Certains connectés au Web enrichissent la base de documents. Un client Unity permet la visualisation des documents dans un espace 3D pour effectuer des requêtes sous la forme d'interaction simple. Ces services travaillent en simultanée pour effectuer des recherches, sauvegarder l'évolution des données et envoyer de l'information sous forme de tableaux et d'interfaces dynamiques.

Afin d'optimiser les résultats obtenus, l'utilisateur doit procéder au paramètrage des différents services et composants. Chacun de ses paramètres, dont l'impact sur les phases de traitement est différent, doit être ajusté pour chercher à obtenir un résultat pertinent. Dans le chapitre suivant, nous étudions l'impact de ces différents paramètres sur plusieurs jeux de données. Ces expérimentations mettent en œuvre des cas proches du réel.

TABLE *3.4* – ***Détail des paramètres de chaque composant et service utilisés dans la chaine de traitement du logiciel WILD. Pour chacun d'eux, nous donnons une description de leur objectif, des paramètres ajustables par l'utilisateur, leur importance et la phase de la chaine de traitement dans lesquelles ils interviennent.***

Service / Composant	Objectif	Paramètres	Importances des paramètres	Phase de la chaine de traitement		
				Détection des signaux faibles potentiels	Enrichissement des thèmes par l'ajout de mots-clés supplémentaires provenant des résultats de recherche	Suivi historique des signaux faibles potentiels
Logiciel WILD	Plateforme d'investigation	1. Nombre de fils d'exécution (threads) [instance]	1. +			
LDA	Méthode de modélisation thématique de documents. Ce modèle reflètent les thèmes sous-jacents dans le contexte global d'un corpus de documents. Utilisé dans : ***LDA*** + ***Word2Vec*** + **Algorithme** et **Attraction / Répulsion**	1. Valeurs de *LDA* choisies	1. +++	X		
Word2Vec	Méthode de plongement de mots. Ce modèle réprésente les mots selon leurs contextes de voisinage dans les documents. Utilisé dans : ***LDA*** + ***Word2Vec*** + **Algorithme**	1. Choix de la base d'apprentissage 2. Taille du vecteur caractéristique	1. +++ 2. ++	X		
w2vSim	Indicateur de cohérence locale. Il repose sur la méthode de plongement de mots, *Word2Vec* et permet de qualifiée la similitude sémantique intrinsèque d'un ensemble de mots (thèmes). Utilisé dans : ***LDA*** + ***Word2Vec*** + **Algorithme**	1. Nombre de mots utilisés	1. ++	X		

Suite à la page suivante

TABLE 3.4 – suite de la page précédente

Service / Composant	Objectif	Paramètres	Importances des paramètres	Phase de la chaine de traitement		
				Détection des signaux faibles potentiels	Enrichissement des thèmes par l'ajout de mots-clés supplémentaires provenant des résultats de recherche	Suivi historique des signaux faibles potentiels
Distance de Bhattacharyya	Indicateur de ressemblance inter-thèmes. Il calcule une distance représentant un lien de ressemblance entre les thèmes de niveaux adjacents. Utilisé dans : ***LDA*** + ***Word2Vec*** + **Algorithme**	1. Seuil de conservation des liens entre thèmes	1. +++	X		
Modélisation thématique multi-niveaux	Algorithme combinant les méthodes de modélisation thématique et de plongement de mots. Utilisé dans : ***LDA*** + ***Word2Vec*** + **Algorithme**			X		
tf-idf	Valeur de pondération des mots selon la base de documents. Il peut être modifié selon les caractéristiques recherchées sur les mots. Utilisé dans : ***LDA*** + ***Word2Vec*** + **Algorithme** et **Attraction / Répulsion**	1. Choix du calcul en fonction des caractéristiques recherchées	1. +++	X	X	X
Composant SMA associé aux documents pour attraction/répulsion	Définit un système multi-agents où des agents "document" évoluant dans un espace 3D sont animés par des forces d'attraction / répulsion. Utilisé dans : **Attraction / Répulsion**	1. Fréquence d'actualisation du système	1. +		X	

Suite à la page suivante

TABLE 3.4 – suite de la page précédente

Service / Composant	Objectif	Paramètres	Importances des paramètres	Phase de la chaine de traitement		
				Détection des signaux faibles potentiels	Enrichissement des thèmes par l'ajout de mots-clés supplémentaires provenant des résultats de recherche	Suivi historique des signaux faibles potentiels
Agent "document"	Représentation d'un document sous la forme d'un agent, défini dans un espace à N dimensions (nombre de thèmes), représenté dans un espace 3D et animé par des forces selon le degré de similarité avec d'autres agents. Utilisé dans : **Attraction / Répulsion**	1. Coefficient d'inertie 2. Coefficient de force 3. Probabilité de séléctionner un agent requête 4. Nombre d'itérations avant changement d'agent	1. + 2. + 3. + 4. +		X	
Composant SMA associé aux mots pour attraction/répulsion	Définit un système multi-agents où des agents "mot" évoluant dans un espace 3D sont animés par des forces d'attraction / répulsion. Utilisé dans : **Attraction / Répulsion**	1. Fréquence d'actualisation du système	1. +		X	
Agent "Mot"	Représentation d'un mot sous la forme d'un agent, défini dans un espace à N dimensions (nombre de mots dans son voisinage), représenté dans un espace 3D et animé par des forces selon le degré de similarité avec d'autres agents. Utilisé dans : **Attraction / Répulsion**	1. Coefficient d'inertie 2. Coefficient de force 3. Probabilité de séléctionner un agent requête 4. Nombre d'itérations avant changement d'agent	1. + 2. + 3. + 4. +		X	
Composant SMA pour la recherche Web	Définit un système multi-agents gérant des agents de recherche. Utilisé dans : **Extraction** et **Recherche**	1. Fréquence d'actualisation du système	1. +		X	X

Suite à la page suivante

TABLE 3.4 – suite de la page précédente

Service / Composant	Objectif	Paramètres	Importances des paramètres	Phase de la chaine de traitement		
				Détection des signaux faibles potentiels	Enrichissement des thèmes par l'ajout de mots-clés supplémentaires provenant des résultats de recherche	Suivi historique des signaux faibles potentiels
Agent de recherche	Agent caractérisé par un groupe de mots permettant de réaliser des requêtes fournies au WebSearchService. Il traite ensuite les résultats retournés pour récupérer les pages Web et les transformer en documents. Utilisé dans : **Extraction** et **Recherche**	1. Nombre de mots voisins récupérés 2. Nombre de résultats de requêtes à récupérer	1. ++ 2. ++		X	X
ProcessIterationService	Service orchestrateur planifiant les exécution de la phase de modélisation thématiques multi-niveaux. La planification est programmée au moyen d'une expression cron. Utilisé dans : ***LDA*** + ***Word2Vec*** + **Algorithme**	1. Expression cron pour la périodicité de la tâche	1. +		X	
SearchAgentService	Service en charge de la création des agents de recherche et produisant les mots utilisés dans les requêtes Web. Utilisé dans : **Extraction**	1. Nombre d'agents par thème 2. Nombre de mots par requête 3. Nombre de requête avant oubli du meilleur agent	1. + 2. ++ 3. ++		X	
WebSearchService	Service lançant les requêtes Web à intervalle regulier pour les agents de recherche. Il passe les requêtes au travers du service TOR et fournit les résultats aux agents. Utilisé dans : **Recherche**	1. Durée d'attente entre les requêtes Web 2. Liste des moteurs de recherche	1. + 2. ++		X	X

Suite à la page suivante

TABLE 3.4 – suite de la page précédente

Service / Composant	Objectif	Paramètres	Importances des paramètres	Phase de la chaine de traitement		
				Détection des signaux faibles potentiels	Enrichissement des thèmes par l'ajout de mots-clés supplémentaires provenant des résultats de recherche	Suivi historique des signaux faibles potentiels
Composant de suivi historique des SMA associés aux mots	Composant de sauvegarde des thèmes permettant leur suivi durant l'investigation. Le composant crée des agents de recherche utilisant des groupes de mots et sauvegarde les résultats sur l'évolution des thèmes. Utilisé dans : **Extraction**	1. Nombre d'agents par thème 2. Nombre de mots récupérés par thème	1. + 2. ++			X
SocketService	Service en charge des échanges d'information entre un client Unity et le modèle que le client visualise. Utilisé dans : **Visualisation**				X	
SocketNewClientService	Service de connexion pour les clients Unity. Il prend en charge les premiers échanges d'information pour faire suivre la connexion au SocketService correspondant. Utilisé dans : **Visualisation**	1. Numéro de port d'accès	1. +		X	
DocumentRESTService	Service d'accès Web. L'utilisateur peut visualiser un ensemble de métrique depuis un navigateur internet en se connectant au service d'interface Web. Utilisé dans : **Visualisation**	1. Numéro de port d'accès	1. +		X	

Chapitre 4

Expérimentations

Sommaire

3.1 Introduction 100
3.2 Etat de l'art 101
3.2.1 Extraction de connaissances 101
3.2.2 Système Multi-Agents 107
3.2.3 *Agent mining* : *Data mining* et Système multi-agents 115
3.3 Système multi-agents et *data mining* proposé 115
3.3.1 Chaine de traitement 116
3.3.2 Système multi-agents associé aux documents 116
3.3.3 Système multi-agents de recherche 123
3.3.4 Système multi-agents associé aux mots 127
3.3.5 Analyse de l'évolution des thèmes 128
3.4 Présentation de l'architecture du logiciel WILD 129
3.4.1 Composants et services du logiciel WILD 131
3.4.2 Composants et services du système d'investigation 134
3.5 Critères et paramétrage de la chaine de traitement 138
3.6 Conclusion 142

4.1 Introduction

L'ensemble de la chaine de traitement d'investigation a été présenté aux chapitres 2 et 3. Celle-ci est intégrée en un logiciel WILD qui propose plusieurs interfaces de gestion. Elle est constituée de deux fonctionalités principales : 1) une modélisation thématique multi-niveaux s'appuyant sur le modèle génératif probabiliste *LDA* et sur le modèle de plongement lexical *Word2Vec* ; 2) un système de recherche de documents pour l'augmentation du corpus initial

et un suivi temporel thématique via un système d'*agent mining*. Ce chapitre 4 se présente comme une étude de pertinence de l'utilisation de cette chaine d'investigation sur 2 corpus de documents : celui des projets H2020, et la base d'articles nous ayant permis d'étudier l'état de l'art sur les *signaux faibles*. Nous proposons ici d'étudier les tendances thématiques présentes dans ces documents en fonction du temps.

4.2 Projets H2020 [2014-2021]

Dans cette étude, le corpus est constitué des résumés des projets H2020 sur la période 2014-2021 provenant de l'Open Data du service d'information sur la recherche et le développement (CORDIS) de l'Union européenne (dump du 06/05/2020). Les sujets sont regroupés par année, l'étude portera donc sur 8 années. Nous utilisons les premières années comme corpus initial, puis nous étudions l'évolution des thèmes et des mots-clés associés sur les années suivantes.

Le site d'information de CORDIS présente 11 grands domaines de news portant sur les projets H2020 (c.f. Figure 4.1).

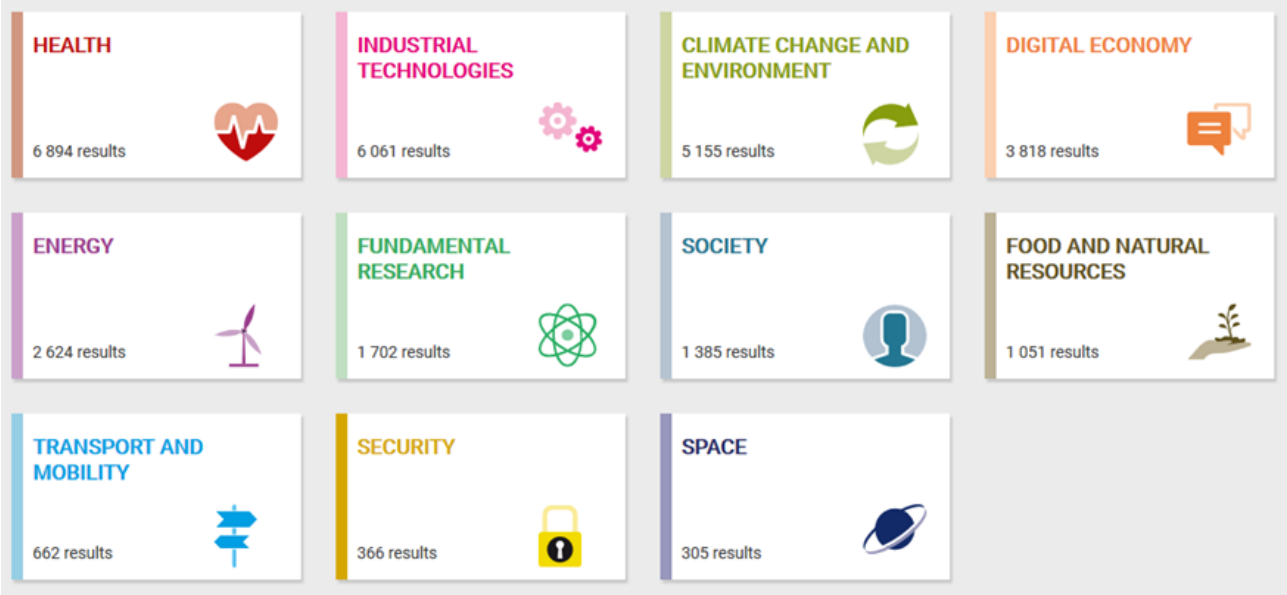


FIGURE *4.1* – ***Le site d'information CORDIS regroupe les news portant sur les projets H2020 en 11 grands domaines (Thèmes) (en date du 09/07/2020).***

4.2.1 Corpus

Pour réaliser le test, nous disposons plus précisement d'un corpus de documents que nous définissons de la façon suivante :

— Corpus initial (années 2014 et 2015) : 5 085 documents

— Index h2020-2016 (année 2016) : 4 943 documents
— Index h2020-2017 (années 2016 et 2017) : 9 912 documents
— Index h2020-2018 (années 2016, 2017 et 2018) : 14 990 documents
— Index h2020-2019 (années 2016, 2017, 2018 et 2019) : 20 516 documents
— Index h2020-2020 (années 2016, 2017, 2018, 2019, 2020 et 2021) : 23 060 documents

Le corpus initial est utilisé lors de la première étape de la chaine de traitement. Avec les thèmes obtenus, nous effectuons des recherches dans chacun des index présentés ci-dessus afin d'étudier l'évolution de différents indicateurs et déterminer si un thème détecté comme *signal faible* potentiel lors de la première étape de la chaine de traitement est effectivement un *signal faible*. Chaque index h2020-année contient les documents de l'année référencée ainsi que l'ensemble des documents des années précédentes exceptées les années 2014 et 2015. Les documents de l'année 2021 sont regroupés avec ceux de l'année 2020 du fait de leur faible nombre.

4.2.2 Résultats

Pour détecter un *signal faible*, il est nécessaire de faire une étude en 2 parties. La première s'applique sur le corpus initial. Celle-ci correspond à l'instant T_0 dans le système et est représentée sur les figures 4.2 et 4.3 par l'encadré "Statique". Le tableau 4.1 permet d'interpréter les thématiques relevées comme pouvant être associées à un *signal faible*. Puis une étude dynamique portant sur l'évolution des thèmes est effectuée afin d'évaluer si chaque thème relève d'un *signal faible*.

4.2.2.1 Partie statique ; étude du corpus initial par la recherche de *signaux faibles* potentiels

L'approche conjointe *LDA/Word2Vec* a été appliquée en faisant varier k sur un intervalle allant de 2 à 8. Le seuil pour la construction de l'arborescence est fixé à 0.6. Nos tests utilisent la méthode de pondération *tf-idf* présentée à la section 2.7.2.2 pour extraire les mots-clés pertinents. Dans notre définition, le *signal faible* est notamment caractérisé par une collection de mots présents de nombreuses fois dans peu de document. Cette propriété peut-être évaluée au moyen de la valeur de *tf-idf*. Nous avons, comme lors des tests de la section 2.7.2.2, calculé cette valeur sur les mots de chaque thème, puis conservé les 20 premiers mots triés par ce même indicateur.

Le tableau 4.1 montre les résultats obtenus sur le corpus H2020 initial, années 2014 et 2015, après élagage, 7 thèmes sont retenus. Afin de découvrir la sémantique des thèmes détectés, nous utilisons 1) les mots-clés classés selon les poids *LDA*, 2) les mots-clés classés selon leur valeur *tf-idf*, et une analyse manuelle des documents couverts par ces mots-clés. Les premiers thèmes

TABLE *4.1* – ***Expérimentation effectuée sur le corpus de documents relatifs aux projets H2020 pour les années 2014 et 2015. Présentation des résultats obtenus après construction de l'arbre et élagage. Pour chaque thème, est indiquée la valeur de*** k ***du*** LDA ***où il est détecté, la cohérence*** Word2Vec***, le nombre de documents associés ainsi que les 20 premiers mots triés par*** tf-idf ***et par*** LDA***.***

Nom du thème	thème 1				thème 2			
Nombre de documents	24				20			
Cohérence du thème (I_1)	141				138			
LDA $k =$	5				4			
	TF-IDF		LDA		TF-IDF		LDA	
Premiers mots du thème	bpm	0.38791	data	0.01309	pcd	0.36096	cells	0.00646
	mmt	0.34551	project	0.00774	rpc	0.34250	cell	0.00637
	blindshell	0.31608	systems	0.00569	atrx	0.31804	clinical	0.00610
	actris-2	0.30479	market	0.00555	g4s	0.31076	cancer	0.00535
	socket	0.30185	technology	0.00475	fld	0.31033	patients	0.00529
	rcms	0.29176	platform	0.00469	cct	0.31033	disease	0.00494
	revault	0.28447	based	0.00467	ccc	0.30380	project	0.00438
	kconnect	0.27618	software	0.00462	auxin	0.30032	development	0.00422
	maya	0.27529	business	0.00427	medulloblastoma	0.28929	molecular	0.00403
	wear3d	0.27309	applications	0.00427	moz	0.28620	human	0.00388
	n400	0.26669	design	0.00393	delirium	0.28212	treatment	0.00383
	body-based	0.26518	services	0.00388	20s	0.27980	health	0.00366
	netik	0.25992	solution	0.00368	apc/c	0.27930	mechanisms	0.00363
	flysec	0.25749	users	0.00367	sorcs1	0.27529	study	0.00338
	tpis	0.24523	's	0.00362	epitaxos	0.27449	diseases	0.00329
	swap.com	0.24077	network	0.00361	gpr37	0.27441	understanding	0.00278
	printoo	0.24077	development	0.00344	ctd	0.27265	studies	0.00276
	unbabel	0.23927	user	0.00343	sleep-active	0.27236	brain	0.00274
	cps	0.07708	solutions	0.00336	fodder	0.27092	genetic	0.00272
	hpc	0.07648	technologies	0.00324	otulin	0.26503	proteins	0.00269

Suite à la page suivante

TABLE 4.1 – suite de la page précédente

Nom du thème	thème 3				thème 4			
Nombre de documents	22				20			
Cohérence du thème (I_1)	131				130			
LDA $k =$	8				5			
	TF-IDF		LDA		TF-IDF		LDA	
Premiers mots du thème	hg	0.45505	models	0.00604	wattsup	0.43486	energy	0.01298
	cct	0.31033	theory	0.00596	ln2	0.31222	market	0.01293
	n400	0.26669	project	0.00584	sofi	0.30753	technology	0.00974
	pge	0.26669	data	0.00573	healex	0.30753	project	0.00974
	body-based	0.26518	understanding	0.00536	agitator	0.30647	production	0.00784
	n-fe2o3	0.25602	methods	0.00486	cortime	0.29655	high	0.00583
	dcops	0.23542	systems	0.00446	payload-lifting	0.29428	process	0.00564
	mwmi	0.22384	model	0.00439	digistone	0.28447	cost	0.00491
	eustace	0.21812	study	0.00420	mirrorpv	0.28361	business	0.00463
	bha	0.21697	climate	0.00417	epitaxos	0.27449	industry	0.00456
	symplectic	0.20818	physics	0.00355	fodder	0.27092	product	0.00438
	emcs	0.20815	approach	0.00326	aeroengine	0.26879	phase	0.00428
	eu-beads	0.19732	problems	0.00326	ngr	0.26340	development	0.00418
	unicon	0.19469	time	0.00325	netik	0.25992	industrial	0.00408
	explicate	0.18552	develop	0.00324	saffron	0.25939	products	0.00397
	normas	0.18495	theoretical	0.00313	n-fe2o3	0.25602	innovative	0.00390
	macro-finance	0.18381	techniques	0.00306	seagate	0.25570	water	0.00372
	weddell	0.18103	field	0.00306	htp	0.25324	developed	0.00362
	mind-wandering	0.17939	computational	0.00301	isoprene	0.25286	manufacturing	0.00361
	rydberg	0.13701	change	0.00296	adex	0.25100	based	0.00359

Suite à la page suivante

TABLE 4.1 – suite de la page précédente

Nom du thème	thème 5				thème 6			
Nombre de documents	19				21			
Cohérence du thème (I_1)	128				123			
LDA $k =$	8				8			
	TF-IDF		LDA		TF-IDF		LDA	
Premiers mots du thème	bpm	0.38791	project	0.01256	nott-300	0.25861	materials	0.01056
	ehri	0.28533	training	0.00992	livox	0.24500	project	0.00888
	mnes	0.27889	social	0.00894	silyliumylidene	0.24306	quantum	0.00711
	tdm	0.27006	researchers	0.00888	iicn	0.23706	applications	0.00623
	ast-fcs	0.25100	science	0.00646	qh	0.22607	properties	0.00568
	penal	0.25027	european	0.00633	oligosaccharide	0.22458	optical	0.00540
	q-tales	0.24306	's	0.00544	lensless	0.22095	high	0.00536
	conscience	0.24077	skills	0.00466	co2-fixation	0.21469	devices	0.00515
	printoo	0.24077	scientific	0.00437	mir	0.21341	technology	0.00515
	gossip	0.23706	europe	0.00433	nwsi	0.20942	based	0.00475
	holocaust	0.23345	public	0.00430	mesocrystal	0.20564	systems	0.00457
	abwet	0.22834	knowledge	0.00424	hydroacylation	0.20319	development	0.00440
	hatha	0.22008	cultural	0.00406	co-pilot	0.19781	light	0.00392
	aveiro	0.22008	study	0.00397	cryo-em	0.19619	design	0.00371
	biopol	0.21802	development	0.00388	sco	0.19396	approach	0.00352
	yoga	0.21570	people	0.00372	cigs	0.19333	chemical	0.00335
	honest	0.21537	understanding	0.00356	electrolyzer	0.19308	field	0.00325
	e-textile	0.20961	studies	0.00340	orr	0.19208	develop	0.00317
	ecec	0.20942	academic	0.00334	rfb	0.18552	control	0.00313
	fiscal	0.11818	analysis	0.00334	thz	0.07579	laser	0.00304

Suite à la page suivante

TABLE 4.1 – suite de la page précédente

Nom du thème	thème 7				
Nombre de documents	20				
Cohérence du thème (I_1)	121				
LDA $k =$	8				
	TF-IDF		LDA		
Premiers mots du thème	thuringia	0.50205	innovation	0.01843	
	zep	0.39693	project	0.01435	
	piano	0.32103	european	0.01285	
	actris-2	0.30479	management	0.01030	
	sdin	0.27708	support	0.00965	
	tdm	0.27006	services	0.00920	
	sc6	0.25783	smes	0.00809	
	prosme	0.25217	europe	0.00757	
	glopid-r	0.24558	activities	0.00706	
	snetp	0.23788	eu	0.00567	
	innocreate	0.22987	sme	0.00523	
	infrafrontier	0.22834	development	0.00498	
	euraxess	0.22757	capacity	0.00492	
	atlantos	0.22384	network	0.00488	
	eccsel	0.22166	key	0.00470	
	rescoop	0.21976	partners	0.00467	
	mecise	0.21976	stakeholders	0.00461	
	regie	0.21882	knowledge	0.00426	
	cca	0.21812	policy	0.00424	
	drr	0.21812	implementation	0.00421	

sont relatifs à des Thèmes liés aux interactions entre l'humain et la machine ou au domaine bio médical et santé. Les mots courts et inintelligibles sont des acronymes. Il est plus difficile, par contre, de distinguer un *signal faible* parmi les thèmes révélés. Nous étudions l'évolution de ces thèmes pour vérifier s'ils correspondent ou non à la définition du *signal faible* décrite à la section 1.3.

La modélisation thématique multi-niveaux permet d'obtenir des thèmes bien construits et homogènes. Les mots-clés qui possèdent un *tf-idf* fort sont bien répartis sur l'ensemble des thèmes. Ces thèmes sont sur des domaines unitaires puisqu'ils ont peu de documents en commun (cf. Tableau 4.2) :

— thème 1. Cette thématique est relative à de nouvelles interfaces homme-machine, recherches technologiques et de solutions de différents secteurs tels que le cloud, les protocoles de communication et les outils de communication ;
— thème 2. Sémantique autour de la santé et du médical avec des mots-clés du domaine de la biologie cellulaire ;
— thème 3. Les mots présents dans ce thème sont relatifs aux modèles théoriques, aux théories de la physique, aux méthodes, aux développement de techniques ainsi qu'aux climats ;
— thème 4. Cette thématique regroupe différents mots-clés de domaines portant sur l'énergie, la production, le développement et l'environnement. Ces derniers sont traités selon les points de vue industriel et commercial ;
— thème 5. La sémantique abordée dans ce thème est associée à la recherche scientifique européenne et universitaire ;
— thème 6. Domaine relatif aux nouveaux matériaux et appareils de mesure ;
— thème 7. Les mots présents dans ce thème sont associés à la recherche en réseaux (collaborations) dans des projets européens.

	thème 1	**thème 2**	**thème 3**	**thème 4**	**thème 5**	**thème 6**	**thème 7**
thème 1	17	0	2	1	3	0	1
thème 2	0	17	1	2	0	0	0
thème 3	2	1	18	1	0	0	0
thème 4	1	2	1	15	1	0	0
thème 5	3	0	0	1	14	0	1
thème 6	0	0	0	0	0	21	0
thème 7	1	0	0	0	1	0	18

TABLE *4.2* – ***Nombre de documents communs entre thèmes.***

Les thèmes possèdent des coefficients de cohérence quasiment équivalents (cf. Tableau 4.1). Dans cette expérimentation, les résultats obtenus ne permettent pas de discerner de thème particulier. Tous les thèmes sont des *signaux faibles* potentiels, aucun thème ne se détache des autres. Ils sont associés à un même nombre de documents. Chacun de ses thèmes peut contenir plusieurs *signaux faibles*, les mots-clés étant les points saillants de ces derniers.

Dès la première itération, l'approche conjointe proposée détecte les thèmes supports de *signaux faibles* potentiels. De ceci, on peut extraire des mots-clés, indicateurs de ces *signaux faibles* potentiels que nous utilisons lors de la deuxième phase de traitement. Dans la suite de notre expérimentation, nous verrons qu'ils sont répartis de manière assez homogène sur les thèmes.

4.2.2.2 Partie dynamique ; évolution des indicateurs de *signaux faibles* potentiels

Nous mettons en œuvre la solution d'*agent mining* proposée dans le chapitre 3. Nous étudions l'évolution des thèmes à partir des recherches effectuées successivement sur les index h2020-année.

La figure 4.2 et le tableau 4.3 montrent l'évolution du nombre de documents récupérés. Les index sont cumulatifs : autrement dit, les documents récupérés sur une année/index pour un thème donné sont également comptabilisés dans les résultats de l'année suivante puisque encore présents dans cet index. Les résultats obtenus montrent que 2 thèmes se démarquent. Les mots-clés présents dans les thèmes 3 et 6 ont permis d'obtenir un nombre important et croissant de documents, principalement sur les périodes 2016-2019.

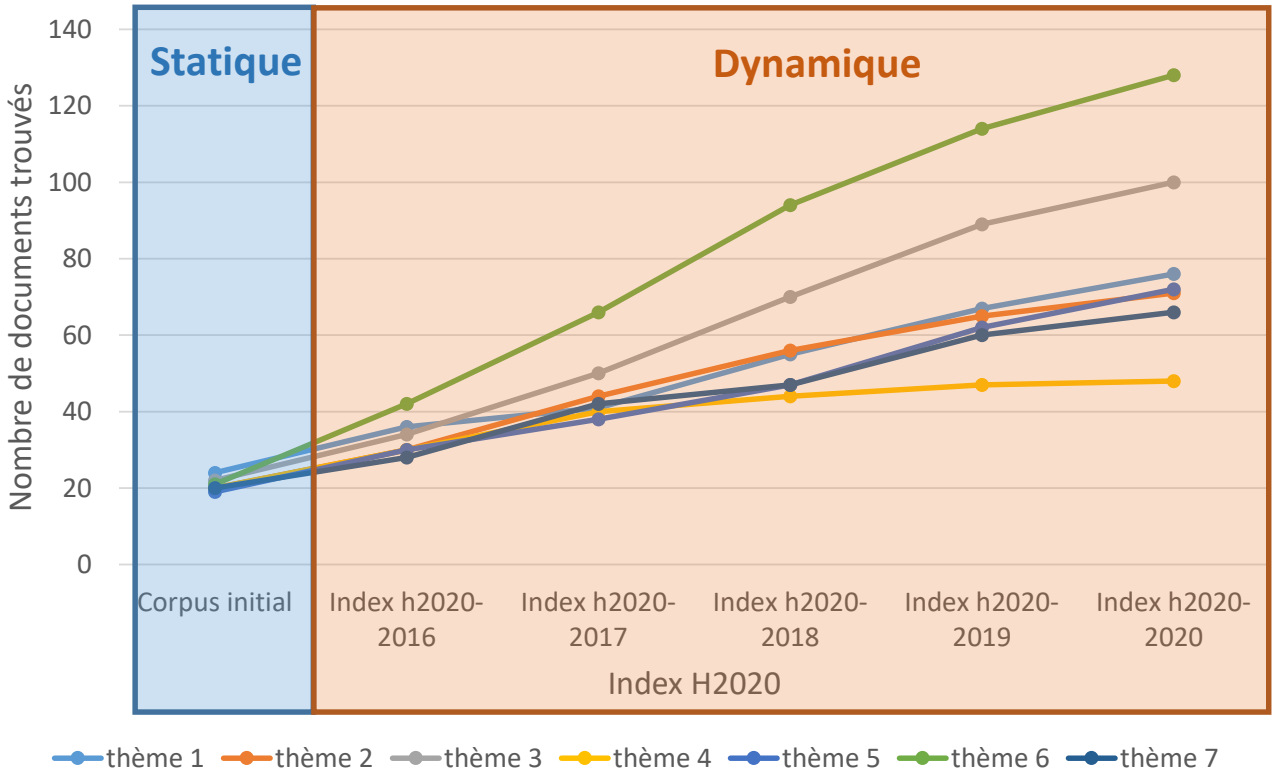


FIGURE *4.2* – ***Evolution du nombre de documents rapportés par les mots-clés de chaque thème. Le corpus initial regroupe les articles des années 2014-2015. Nous indiquons le nombre de documents rapportés qui contiennent au moins un mot parmi les 20 premiers mots-clés, et ceci sur chacun des index.***

	Corpus initial	Index h2020-2016	Index h2020-2017	Index h2020-2018	Index h2020-2019	Index h2020-2020	Rapport fin/initial
thème 1	24	36	41	55	67	76	317%
thème 2	20	30	44	56	65	71	355%
thème 3	22	34	50	70	89	100	455%
thème 4	20	30	40	44	47	48	240%
thème 5	19	30	38	47	62	72	379%
thème 6	21	42	66	94	114	128	610%
thème 7	20	28	42	47	60	66	330%

TABLE *4.3* – ***Tableau montrant le nombre de documents rapportés par les mots-clés de chaque thème. Une colonne détaille le rapport du nombre de documents rapportés à partir du dernier index sur le nombre de documents du corpus initial pour chaque thème.***

La tendance des 2 thèmes se confirme dans la figure 4.3 où pour chaque thème, et pour chaque index, la somme des valeurs *tf-idf* est calculée. Les thèmes 3 et 6 ont les valeurs les plus basses. Le nombre de documents rapportés étant plus important pour ces thèmes, leurs valeurs *tf-idf* baissent notablement.

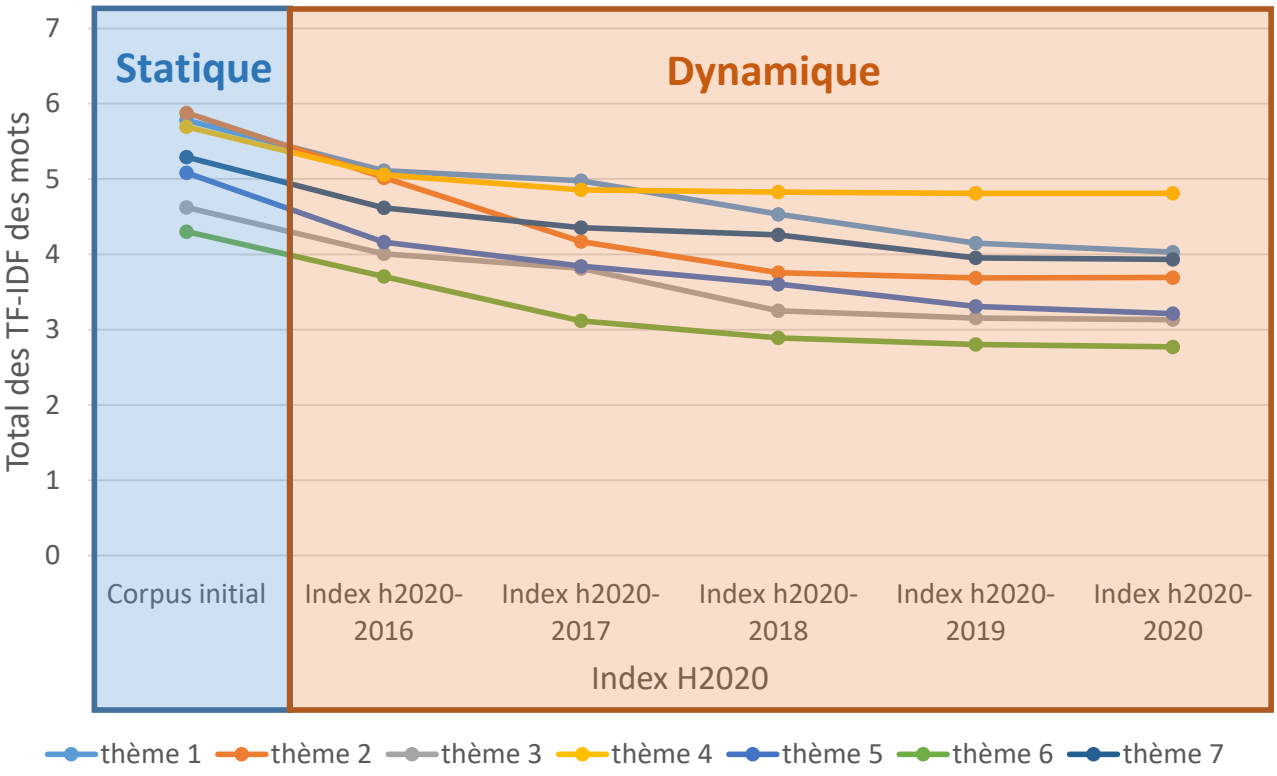


FIGURE *4.3* – ***Evolution de la valeur de*** tf-idf ***pour chaque thème. Le corpus initial regroupe les articles des années 2014-2015. La valeur est calculée sur les 20 premiers mots de chaque thème après qu'ils aient été triés.***

Les 7 thèmes, détectés comme des *signaux faibles* potentiels sur le corpus inital sont associés

à des mots-clés, indicateurs de ces *signaux faibles*, qui permettent donc un apport important de documents portants sur ces thèmes. Ceci est encore plus notable pour les thèmes 3 et 6.

Si l'on regarde maintenant le nombre de documents rapportés par les mots-clés les plus pertinents (cf. Tableau 4.4 et Figure 4.4), on peut alors dégager sur chaque période les domaines ciblés des projets H2020 relatifs aux thèmes *signaux faibles* potentiels initiaux.

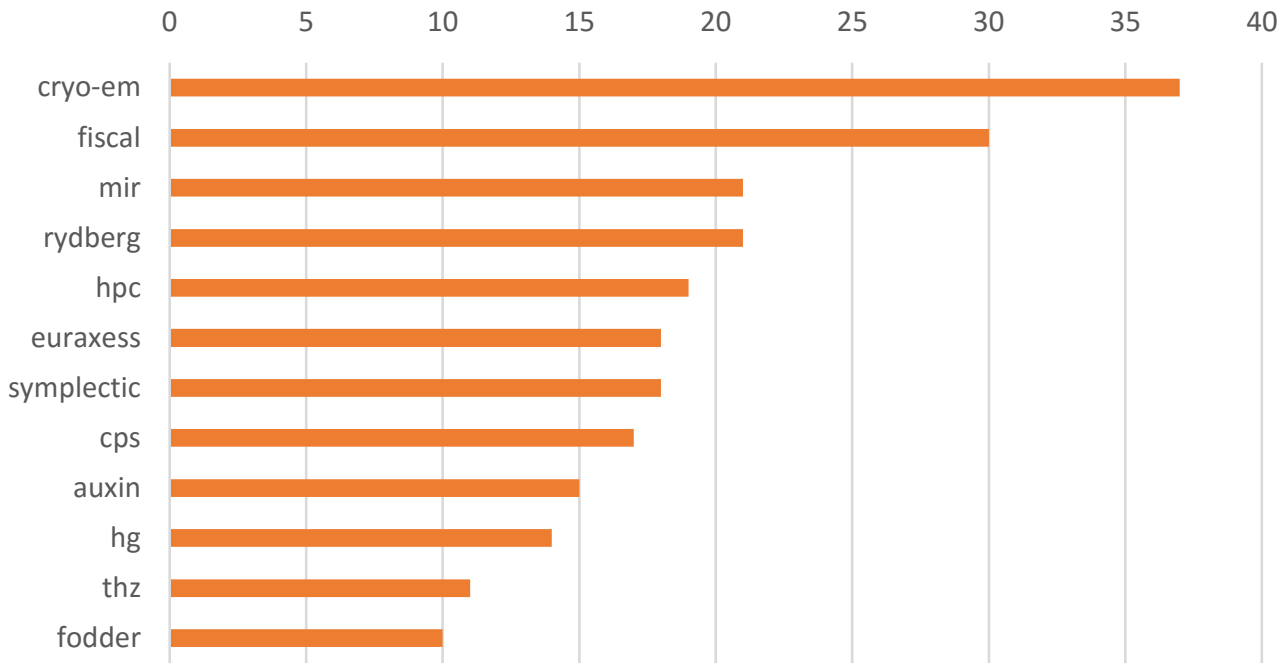


FIGURE *4.4* – ***Représentation graphique du nombre de documents obtenus après requêtes des mots dans un moteur de recherche. Seuls les mots rapportant un nombre de documents significatifs sont présents.***

Parmi ces mots-clés, nous trouvons :

- thème 1
  - hpc : calcul haute performance permettant de traiter les données et d'effectuer des calculs complexes à des vitesses élevées.
  - cps : système cyber-physique où des éléments informatiques collaborent pour le contrôle et la commande d'entités physiques. Il est utilisé dans des innovations technologiques.
- thème 2
  - auxin : hormone de plantes ; elle est utilisée dans des études d'impact de changement d'environnement et de température visant à déterminer ses possibilités d'adaptations.
  - fodder : fourrage utilisé en agriculture. Ce mot-clé est utilisé dans des projets autour de l'optimisation dans la production de produits finaux tels que le lait, la viande. Il fait l'objet d'étude de marché.
- thème 3
  - hg : symbole chimique du mercure. Il fait l'objet d'étude d'impact physique sur

Mots	Nb document(s)	Mots	Nb document(s)
cryo-em	37	sc6	3
fiscal	30	20s	3
mir	21	ecec	2
rydberg	21	cct	2
hpc	19	qh	2
euraxess	18	weddell	2
symplectic	18	aveiro	2
cps	17	rcms	2
auxin	15	zep	2
hg	14	sofi	2
thz	11	mnes	2
fodder	10	pge	2
explicate	8	sleep-active	2
rfb	8	mmt	2
pcd	8	delirium	2
holocaust	8	bpm	2
cigs	8	mind-wandering	2
htp	7	g4s	2
honest	6	gossip	2
electrolyzer	6	rescoop	1
cca	5	atrx	1
isoprene	5	wear3d	1
orr	5	infrafrontier	1
socket	5	apc/c	1
tdm	5	medulloblastoma	1
penal	4	drr	1
aeroengine	4	actris-2	1
conscience	3	ehri	1
oligosaccharide	3	glopid-r	1
ctd	3	eccsel	1
co-pilot	3	bha	1
sco	3	fld	1
maya	3	snetp	1
thuringia	3	payload-lifting	1
prosme	3	macro-finance	1

TABLE *4.4* – ***Nombre de documents obtenus après requêtes à partir de mots-clés dans un moteur de recherche (ElasticSearch) composé des 23 060 résumés des projets H2020. Ces derniers couvrent la période 2016-2020 et ont été extraits à partir de l'Open Data du service CORDIS de l'union européenne (cf. section 4.2.1)***

l'environnement.

— symplectic : géométrie symplectique étudié dans le cadre de projet en mathématiques fondamentales.

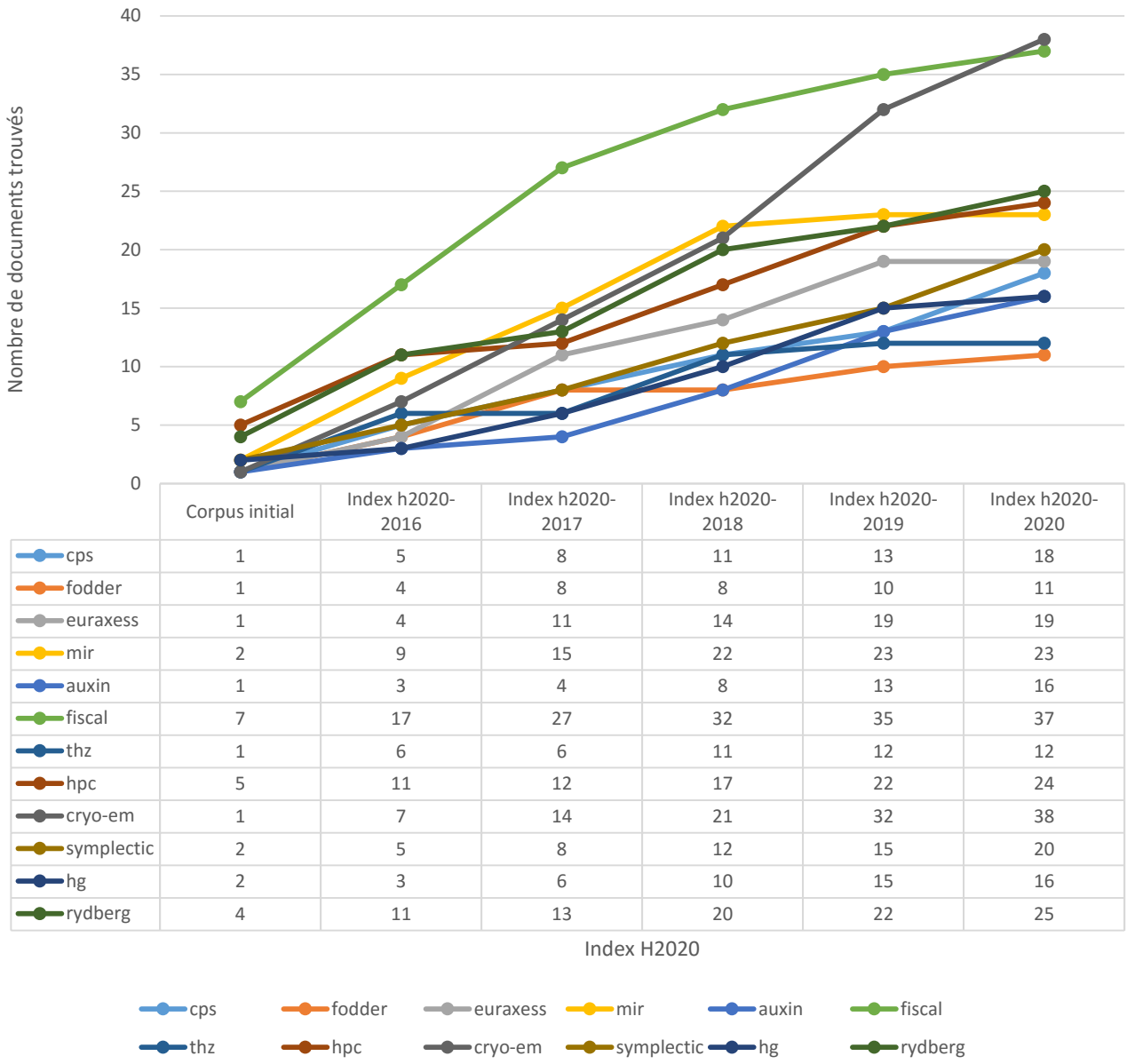
— Rydberg : nom de l'état excité d'un atome en physique atomique. La particularité

de cet atome est sa grande taille.
- thème 4
  - fodder : présenté précedemment
- thème 5
  - fiscal : Etude de politique budgétaire dans divers projets (études de coûts, crises économiques, études historiques, modèles d'union monétaire et fiscale,...)
- thème 6
  - thz : Symbole du térahertz, unité de mesure de fréquence utilisé en électronique (industrie, spectroscopie,...).
  - mir : division du rayonnement optique (Infrarouge moyen) utilisée par des composants électroniques
  - cryo-em : nouvelle technologie d'acquisition liée à la microscopie électronique.
- thème 7
  - euraxess : initiative européenne pour fournir information et support aux chercheurs sur les métiers de la recherche en Europe.

Parmi ses mots-clés, certains se démarquent par leurs évolutions dans les résultats de recherches sur les différents index (cf. Figure 4.5). Des mots-clés tels que "cps", "cryo-em" et "fiscal" rapportent beaucoup de documents. "cryo-em", "thz" et "mir" sont liés au même domaine dans le thème 6. "mir" montre une forte tendance, détecté comme *signal faible* à l'itération 0, son évolution le présente commme un indicateur de *signal faible*. "fiscal" présente une forte évolution lors des premières années qui ralenti sur les dernières années. Il est déjà *signal fort*. D'autres mots comme "rydberg", "cps", "hpc", "symplectic" et "euraxess" sont intéressants aussi bien que moins présents. Leurs évolutions montrent qu'un intérêt existe sur ses domaines. Les mots "auxin", "hg", "fodder" montrent des évolutions moins développées ne permettant pas encore de déduire des *signaux faibles*.

On a détecté lors de la première itération des mots-clés relatifs à un *signal faible* lié à un thème sur le thème de la santé et du médical. Ces mots-clés ont permis de récupérer des documents dans l'index "Index h2020-2016". Lors de la seconde itération, nous relançons la phase de modélisation thématique multi-niveaux, afin d'actualiser les thèmes de notre système multi-agents. Les nouveaux résultats ont permis d'obtenir 8 thèmes. Nous avons donc réussi à faire émerger un thème qui n'était pas initialement présent en tant que tel. Ce nouveau thème est la séparation du thème 2 en deux thèmes, relatifs au secteur médical pour l'un, et à la biologie cellulaire pour l'autre (cf. Tableau 4.5).

L'analyse longitudinale a permis de dissocier deux thèmes qui apparaissaient comme "mélangés" initialement : il s'agit du thème "santé/médical" et du thème "biologie cellulaire".



	Corpus initial	Index h2020-2016	Index h2020-2017	Index h2020-2018	Index h2020-2019	Index h2020-2020
cps	1	5	8	11	13	18
fodder	1	4	8	8	10	11
euraxess	1	4	11	14	19	19
mir	2	9	15	22	23	23
auxin	1	3	4	8	13	16
fiscal	7	17	27	32	35	37
thz	1	6	6	11	12	12
hpc	5	11	12	17	22	24
cryo-em	1	7	14	21	32	38
symplectic	2	5	8	12	15	20
hg	2	3	6	10	15	16
rydberg	4	11	13	20	22	25

FIGURE *4.5* – ***Evolution, pour chaque index, du nombre de documents rapportés par les mots-clés les plus pertinents.***

4.2.3 Conclusion

Les thèmes obtenus au cours de l'analyse statique possèdent une valeur de cohérence quasi identique. Ils sont tous potentiellement supports de *signaux faibles*. Grâce à l'étude sur les 8 années, deux thèmes se détachent tout particulièrement des thèmes 3 et 6 qui représentent en fin de période H2020 des études et recherches très présentes. Ceux-ci sont donc à suivre. Comme on peut le voir sur le tableau 4.6, ces deux thèmes partagent très peu de documents avec les autres thèmes.

On peut considérer que les thèmes 3 et 6 représentaient, en début de période H2020, des domaines de recherche en émergence, les mots-clés de ces thèmes constituant les points saillants

Nom du thème	thème 0				thème 5			
Cohérence du thème (I_1)	142				125			
LDA $k =$	7				7			
	TF-IDF		LDA		TF-IDF		LDA	
Premiers mots du thème	rpc	0.34536	clinical	0.01261	ccc	0.30633	cell	0.01064
	epitaxos	0.27678	patients	0.01119	moz	0.28859	cells	0.01062
	peanut	0.26342	health	0.00927	sorcs1	0.27759	molecular	0.00639
	imuc	0.25989	cancer	0.00797	gpr37	0.27669	mechanisms	0.00583
	biostealth	0.25476	treatment	0.00775	otulin	0.26724	proteins	0.00469
	dkd	0.25145	disease	0.00722	met1-ub	0.26724	development	0.00463
	fgfs	0.23757	project	0.00595	curli	0.26653	human	0.00448
	decubitus	0.23576	medical	0.00540	healthspan	0.26614	genetic	0.00447
	frataxin	0.23530	care	0.00533	hsp70	0.26559	understanding	0.00435
	uncap	0.23118	patient	0.00495	bgt	0.26478	protein	0.00409
	melanocyte	0.22871	diagnostic	0.00479	myopia	0.26076	biology	0.00406
	ap4a	0.22846	risk	0.00473	pi3k	0.25978	role	0.00387
	formac	0.22846	market	0.00464	cfs	0.25434	project	0.00382
	mwmi	0.22571	diseases	0.00429	kef	0.25235	dna	0.00368
	louisa-3d	0.22136	based	0.00401	nmj	0.25145	identify	0.00359
	ths	0.21994	detection	0.00360	cb1	0.24823	function	0.00352
	avf	0.21668	device	0.00355	apol1	0.24586	cellular	0.00337
	idili	0.20861	diagnosis	0.00349	t6ss	0.24071	gene	0.00333
	5-part	0.20771	development	0.00349	duts	0.23684	aim	0.00322
	nafld	0.20663	develop	0.00339	prdm9	0.23363	model	0.00302

TABLE *4.5* – ***Présentation de deux thèmes émergents du thème 2 présent dans le tableau 4.1 obtenus après construction de l'arbre et élagage lors de l'itération suivante. Le corpus contient les documents initiaux ainsi que ceux récupérés dans l'index h2020-2016. Pour chaque thème, est indiquée la valeur de*** k ***du*** LDA ***où il est détecté, la cohérence*** Word2Vec ***ainsi que les 20 premiers mots triés par*** tf-idf ***et par*** LDA.

	thème 1	**thème 2**	**thème 3**	**thème 4**	**thème 5**	**thème 6**	**thème 7**
thème 1	29	0	2	0	20	0	1
thème 2	0	39	2	10	0	0	0
thème 3	2	2	74	0	0	0	0
thème 4	0	10	0	19	0	0	0
thème 5	20	0	0	0	28	0	5
thème 6	0	0	0	0	0	106	1
thème 7	1	0	0	0	5	1	39

TABLE *4.6* – ***Nombre de documents communs entre thèmes après analyse sur les 8 années.***

de ces domaines. Ces thèmes sont unitaires car ne partageant pas de liens avec d'autres thèmes. Nous montrons également qu'à partir de 2016, deux thèmes émergents du thème 2, le premier portant sur le domaine médical, le second sur la biologie cellulaire. Le suivi sur une longue période permet d'évaluer la dynamique de ces domaines.

4.3 Analyse bibliographique d'une base de données d'articles

Comme présenté dans le chapitre 1, nous expérimentons notre chaine de traitement sur la base d'articles étudiés par Mühlroth. Cette étude a pour objectif la recherche de *signaux faibles* potentiels au sein des résumés des articles scientifiques relevés par l'auteur dans ses travaux et leur évolution depuis la publication de l'article [MG18] en Février 2018.

4.3.1 Corpus

Pour ce test, le corpus de base est constitué des 91 articles relevés par Mühlroth. Parmi ceux-ci, nous recherchons un ou plusieurs *signaux faibles* potentiels. Pour la partie *agent mining*, permettant l'enrichissement de la donnée, nous éxécutons la même commande de recherche (que Mühlroth) sur Web of Science. Les articles sur la période d'avril 2017 à septembre 2020 sont alors extraits(cf. Tableau 1.3). Les résumés de ces articles sont placés dans un index Elasticsearch. Cet index simule un moteur de recherche sur lequel nous effectuons des requêtes grâce aux mots-clés obtenus lors de la première phase de la chaine de traitement. Nous disposons donc de deux bases documentaires, l'une appelée base initiale recueille des articles sur la période 1997-2017, l'autre appelée Index ElasticSearch pour les articles datant de la période 2017-2020. L'ensemble des bases documentaires est récapitulé dans le tableau 4.7.

Corpus	Nb de documents
Base initiale	91
Index ElasticSearch	537

TABLE *4.7* – ***Constitution de la base de documents utilisée à l'initialisation de la chaine de traitement (base initale) et ajout d'un index ElasticSearch pour les requêtes Web.***

Dans ses travaux, Mühlroth [MG18] organise les articles selon la classification PEST. Cette classification répartit les documents selon les domaines politique, économique, social et technologique. Le nombre de documents pour chacun de ces domaines n'est pas équitablement réparti. Ainsi, le tableau 4.8 montre que le domaine technologique est majoritaire, les domaines politiques et économiques sont, quant à eux, très peu représentés. Cette répartition a nécessairement un impact sur les résultats.

Cette base documentaire a les spécificités suivantes :

— Certaines thématiques sont sous-représentées dans le corpus. Le domaine des technologies représente 63% des articles relevés par Mühlroth.

— Le nombre de document est limité : 91 articles seulement.

Domaine	Nb de documents
Politique	4
Economique	7
Social	22
Technologique	58
Total	**91**

TABLE *4.8* – ***Répartition du nombre d'articles utilisés dans les travaux de Mühlroth selon les 4 domaines de la classification PEST.***

— Pour les deux périodes, le nombre de mots présents dans chaque document est faible. En effet, les documents retenus sont les résumés des articles dont la taille est entre 1ko et 2ko.

4.3.2 Mise en œuvre

Dans le cadre de cette expérimentation, certains services du logiciel WILD ne sont pas nécessaires. Dans le graphique 4.6, nous présentons seulement les composants et services utilisés pour la réalisation du test.

La chaine de traitement suit une procédure ciblée composée d'étapes bien précises pour permettre la reproductibilité du test. Cet enchainement d'étapes permet d'éviter les inter-blocages des composants et services constituant le logiciel WILD :

1. Au démarrage, le logiciel WILD lancé recherche la présence des fichiers de configuration des instances d'investigation existantes. Une configuration est créée pour chaque nouvelle instance d'investigation.
2. L'instance d'investigation initialise les composants et services qui la composent.
3. Le composant SMA associé aux documents charge l'ensemble des documents présents dans le corpus initial. Pour chacun d'entre eux, il crée un agent "document" qu'il place dans son espace.
4. Le service du processus itératif "ProcessIterationService" exécute la première phase de la chaine de traitement pour obtenir les thèmes présents dans le corpus initial. Il ajoute ensuite les thèmes au composant SMA associé aux mots. Cette étape entraine la mise à jour et sauvegarde des informations de l'instance d'investigation.
5. Une nouvelle instance du composant de suivi historique des SMA associé aux mots est créée afin de sauvegarder les thèmes et ainsi étudier l'évolution des mots qu'elle contient.
6. L'instance d'investigation indique à tous les agents "document" de mettre à jour leur poids en fonction des nouveaux thèmes.
7. Le composant de suivi historique lance des recherches avec des agents de recherche à partir des mots des thèmes sauvegardés.

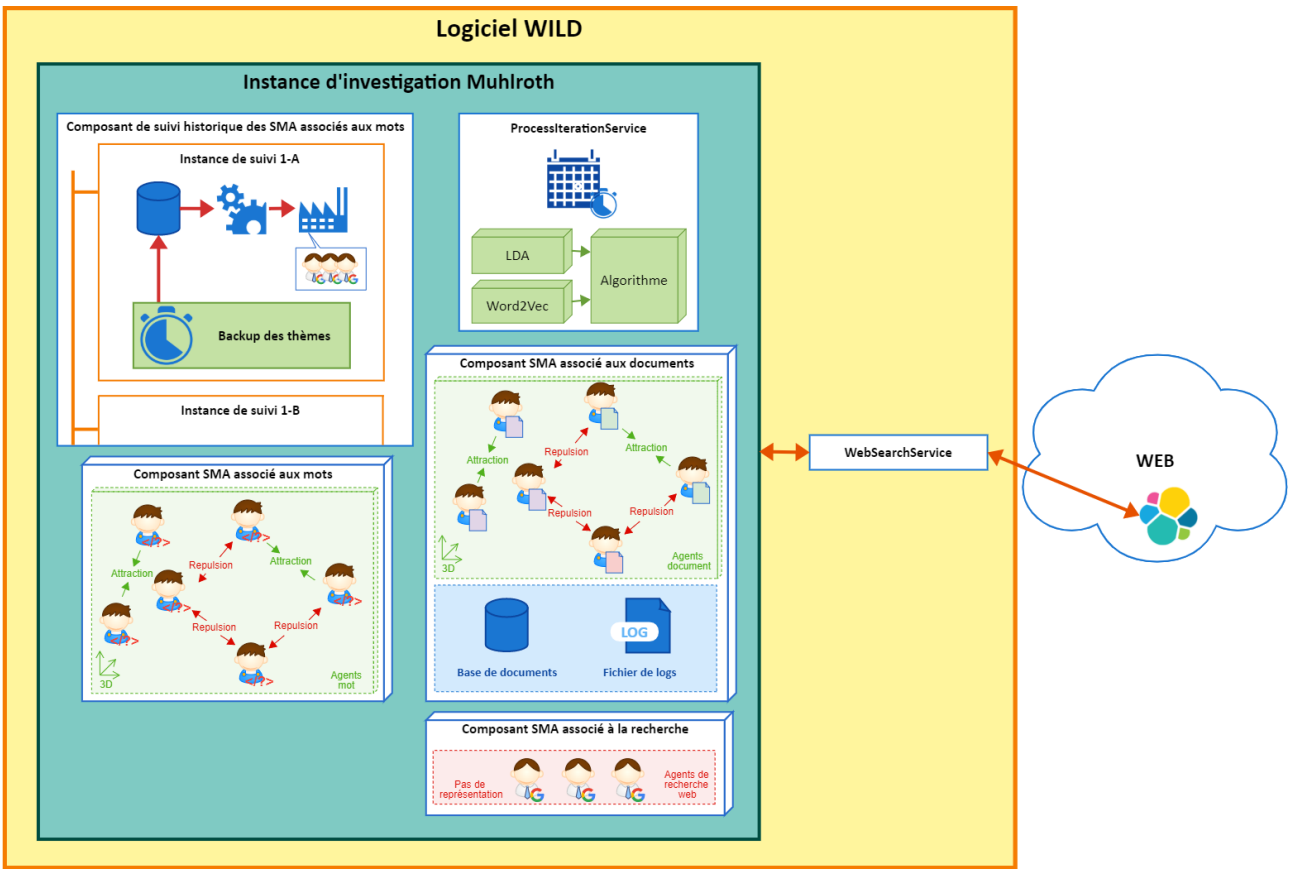


FIGURE *4.6* – ***Composants et services utilisés pour l'expérimentation sur la base documentaire de Mühlroth***

8. Les agents de recherche exécutent les requêtes demandées et ajoutent les nouveaux documents au corpus de l'instance d'investigation.
9. Les composants SMA associés aux documents et aux mots mettent à jour régulièrement les poids des agents "document" et les valeurs *tf-idf* des agents "mot" au fur et à mesure que des nouveaux documents sont ajoutés aux corpus.
10. Quand le composant de suivi historique a effectué toutes les recherches, ce dernier sauvegarde une nouvelle version des thèmes ainsi que les valeurs *tf-idf* des agents "mot" qu'il contient.
11. Une nouvelle itération du processus "ProcessIterationService" est exécuté sur le nouveau corpus composé du corpus initial et des documents récupérés afin d'obtenir les nouveaux thèmes.

La méthode de modélisation thématique multi-niveaux, composée de *LDA/Word2Vec* ainsi que de notre algorithme, a été évaluée en faisant varier k sur un intervalle allant de 2 à 12. Le seuil pour la détermination de l'arborescence est fixé à 0.3. Le calcul pour l'indicateur de cohérence intra-thème est réalisé sur les 40 premiers mots de chacun d'entre eux.

La première phase de chaine de traitement (la modélisation thématique multi-niveaux)

terminée, les mots des thèmes sont triés au moyen d'une méthode de pondération *tf-idf* modifiée (cf. Equation 4.1).

$$\textit{tf-idf}_i = f_{t_i,D_i} \cdot \log \frac{|D|}{|D_i|} \quad \text{où} \quad f_{t_i,D_i} = \frac{1}{|D_i|}(\sum_{j=0}^{D_i} \frac{n_{i,j}}{n_j}) \quad \text{et} \quad D_i = \{d_j : t_i \in d_j\} \tag{4.1}$$

$\textit{tf-idf}_i$ 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_i le nombre de documents où le terme t_i apparaît, n_j le nombre de mots dans le document d_j et $n_{i,j}$ le nombre d'occurrence de t_i dans le document d_j. f_{t_i,D_i} correspond à la moyenne du nombre d'apparition du mot t_i dans tous les documents.

4.3.3 Principaux résultats obtenus

Dans le chapitre 1, nous présentons une méta-analyse bibliographique sur la base d'articles étudiés par Mühlroth. Nous révélons des *signaux faibles* potentiels non référencés comme tel dans ses travaux.

Un thème sur le domaine de la santé détecté sur la période 1997-2017 se confirme sur la période 2017-2020 comme étant un *signal faible* potentiel et un attrait fort dans la littérature depuis les récentes crises sanitaires telles que la pandémie de Covid-19 et l'épidémie de maladie à virus Ebola sur cette période 2017-2020.

Nous détectons deux nouveaux thèmes comme *signaux faibles* potentiels sur la période 2017-2020. Le premier porte sur le domaine environnement/climatologie/météorologie qui est un enjeu sociétal et scientifique récent avec des sujets tels que l'élévation du niveau des mers, les risques d'inondations et des conséquences à l'échelle mondiale. Un second porte sur le domaine des matériaux et leurs impacts environnementaux qui est également important aujourd'hui avec des sujets comme la pollution, l'extraction des métaux rares et leurs utilisations dans l'industrie.

4.3.3.1 Détermination du seuil de détection du *signal faible* sur la base documentaire de Mühlroth

Comme le montre la figure 1.5, sur la période 1997-2017, seuls les prémisses du *signal faible* du domaine "Médical/Santé" ont pu être détectés. Ceci est dû à plusieurs facteurs :

— le nombre de documents tous thèmes confondus ;
— le nombre insuffisant de document, portant sur ce domaine ;
— la taille très réduite des résumés ;
— la nature de ces résumés, très génériques ;

— les thèmes principaux de l'étude bibliographique : détection de *signaux faibles* et *tendances* qui sont des domaines transversaux.

Dans cette nouvelle expérimentation, nous cherchons à déterminer le nombre de documents supplémentaires qu'il est nécessaire d'ajouter pour détecter le domaine "médical et santé" comme *signal faible*. Nous avons choisi empiriquement 6 documents relevant de ce thème que nous avons rajoutés au corpus initial. Le tableau 4.9 montre maintenant l'apparition de ce thème avec les mots-clés associés tels que "health", "drug", "health-related", "clinical", "treatment" quand ils sont triés par les poids *LDA*, et les mots "cognitive", "organ", "therapies", "interventions" quand triés par *tf-idf*.

Nom du thème	thème 1 / 6			
LDA k =	7			
	TF-IDF		LDA	
Premiers mots du thème	sevoflurane	0,15081	health	0,02750
	deqi	0,11999	series	0,01768
	rosc	0,10397	time	0,01768
	cognitive	0,10054	deqi	0,01572
	organ	0,10054	rosc	0,01375
	delirium	0,10054	group	0,01375
	deficiency	0,10054	drug	0,01375
	ad	0,08356	health-related	0,01179
	co2	0,07426	clinical	0,01179
	boards	0,06931	co2	0,00982
	drug	0,06775	conditions	0,00982
	therapies	0,06267	study	0,00982
	health-related	0,06053	studies	0,00982
	interventions	0,06000	items	0,00786
	fifty-four	0,05512	articles	0,00786
	postanesthesia	0,05027	elements	0,00786
	protective	0,05027	term	0,00786
	populations	0,05027	drugs	0,00786
	decades	0,05027	key	0,00786
	anesthetic	0,05027	treatment	0,00786

TABLE *4.9* – ***Détail du thème obtenu par notre approche de modélisation thématique multi-niveaux sur le corpus initial contenant 6 documents supplémentaires portant sur le domaine médical et santé.***

Sur cette bibliographie, particulièrement difficile pour les raisons explicitées précédemment, il est nécessaire d'avoir 14% de documents issus du thème *signal faible* pour que celui-ci soit détecté, soit un taux 3 fois plus important que le taux détecté au chapitre 2 (5%) dans des conditions cependant plus aisées.

Conclusion générale et perspectives

Aspects conception et mise en œuvre

La solution WILD que nous avons conçue pour la recherche de *signaux faibles* à partir d'un corpus de documents repose sur un ensemble de séquences automatisées : traitement de l'information brute, enrichissement de celle-ci au moyen de requêtes web, et présentation des résultats sous la forme de tableaux de bord.

Certaines étapes nécessitent cependant l'intervention d'experts du domaine afin d'interpréter la sémantique des thèmes et les mots-clés saillants relatifs à ces derniers.

La solution WILD est composée des étapes suivantes, dans l'ordre :
— collecte des données ;
— nettoyage et pré-traitement des données ;
— projection et transformation des données par modélisation thématique et plongement de mots ;
— exploration de données par algorithme et évaluation de la pertinence des *signaux faibles* potentiels obtenus ;
— suivi des thèmes *signaux faibles* potentiels par recherche complémentaire.
La construction de cette chaine a été guidée par notre définition du *signal faible* :

Definition. Un *signal faible* est caractérisé par un faible nombre de mots par document et présent 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).

Le logiciel réalise donc une procédure d'enquête dont les fonctions sont : - une analyse semi-automatique de contenus avec un minimum d'*a priori* ; - l'agrégation de connaissances ; - une visualisation analytique.

Les contributions décrites dans les différents chapitres ont donné lieu à des publications dans un journal et des conférences internationales et nationales (cf. chapitre 4.2). Ces contributions sont décrites ci-après et des perspectives sont proposées pour chacune d'elles.

Dans le cadre de ces travaux, nous avons priorisé la construction d'une chaine de traitement complète en étudiant le résultat de chaque bloc qui la compose. De cette manière, nous avons pu réaliser une preuve de concept avec des données synthétiques et réelles. Le logiciel comprenant un ensemble de composants et services, chacun d'eux peut être modifié indépendamment.

Dans les sections suivantes, nous abordons les différentes "contributions" apportées par ces travaux et proposons des "perspectives" d'améliorations possibles.

Approche statique et modélisation thématique multi-niveaux

Contributions. Nous proposons une solution d'analyse semi-automatique de documents pour extraire des *signaux faibles* potentiels. Cette solution repose sur une modélisation thématique multi-niveaux qui combine modélisation thématique *LDA* et plongement de mots avec *Word2Vec*. Notre démarche repose donc sur un ensemble d'algorithmes de la littérature et de critères :

- ***LDA*** permet d'obtenir les mots-clés les plus pertinents pour la sémantique du thème ;
- ***Word2Vec*** donne un critère de cohérence des mots appartenant aux thèmes ;
- **la distance de Bhattacharyya** permet d'obtenir les thèmes les plus disjoints possibles par le calcul des liens de ressemblance entre les thèmes ;
- ***tf-idf*** calcule les mots-clés saillants d'un thème pouvant représenter plusieurs *signaux faibles* potentiels.

Perspectives. Plusieurs paramètres sont ajustables tels que le nombre de niveaux *LDA* composant l'arbre, le nombre de mots pris en compte dans chaque thème pour évaluer leur cohérence, le seuil de l'indicateur de ressemblance reliant les thèmes entre deux niveaux de l'arborescence,... Parmi ceux-ci, la pertinence des thèmes obtenus et groupe de mots qui les composent sont ré-évalués au moyen de la méthode de pondération *tf-idf*. Cette pondération peut être remplacée selon les critères sur lesquels les *signaux faibles* doivent être détectés. Dans l'expérimentation sur les articles de Mühlroth, celle-ci a montré ses limites dans des conditions difficiles comme un faible nombre de documents contenant peu de mots. On pourrait envisager de prendre en compte la pondération *LDA* des mots dans les thèmes pour la méthode de pondération *tf-idf*. Le recours à un expert du domaine permet d'évaluer la pertinence des résultats et d'identifier les thèmes porteurs de *signaux faibles* potentiels.

Approche dynamique et solution d'*agent mining*

Contributions. Le suivi temporel de *signaux faibles* permet de découvrir ceux devenant des *signaux forts*. Pour cela, la corrélation à un contexte d'information plus large effectuée grâce à des requêtes sur des moteurs de recherche enrichit le corpus initial et permet un suivi longitudinal des *signaux faibles* potentiels.

La solution d'*agent mining* que nous proposons, combinaison de *data mining* et de systèmes multi-agents, crée des agents de recherche qui enrichissent le corpus par des recherches Web

à partir des thèmes et mots-clés saillants obtenus durant l'approche statique. Les pages Web, résultats de ces recherches, sont ajoutées sous la forme de nouveaux agents "document". Les documents et mots-clés obtenus durant l'approche statique sont représentés dans un espace 3D sous forme d'agents animés par des forces d'attraction/répulsion. L'utilisateur peut interagir avec les agents en figeant et déplaçant des agents, réorganisant alors l'affichage des autres agents.

Cette approche dynamique permet de suivre l'évolution des mots-clés et les résultats des recherches de documents associés afin de déterminer si un *signal faible* potentiel devient *signal fort*, selon le nombre de documents reçus.

Perspectives. Le nombre de documents obtenus est un critère de suivi temporel mais nous avons besoin de l'avis d'expert du domaine pour déterminer si parmi les thèmes étudiés, il y a effectivement présence ou non de *signaux faibles*.

Donner la possibilité à l'expert de pouvoir interagir avec l'ensemble des agents pour le guider dans sa recherche documentaire n'est présente dans la solution logicielle que sous la forme d'une ébauche. Il est nécessaire, pour évaluer sa pertinence, d'enrichir l'interface, par exemple avec les métadonnées relatives aux documents.

Architecture du logiciel WILD

Contributions. Cette plateforme d'investigation fournit un ensemble de services et composants à destination de différents utilisateurs tels que des lanceurs d'alertes et des journalistes dans un objectif de prise de décision. L'ensemble des services et composants travaillent en parallèle, l'ensemble est définit de manière modulaire. Chaque instance d'investigation dispose de données, documents, mots et logs spécifiques. L'utilisateur doit déterminer par un réglage fin l'ensemble des paramètres du logiciel WILD et des instances d'investigation.

Perspectives. Le logiciel étant modulable pour chaque instance d'investigation, de nouveaux services peuvent être ajoutés. Le service de recherche Web peut founir d'autres moteurs de recherche qu'ElasticSearch et Qwant. La visualisation des agents dans l'espace 3D pourrait se faire dans un autre client (par exemple, une interface Web). Des moyens d'interactions sociales peuvent être rajoutés comme une messagerie instantanée ou un service d'audioconférence dans chaque instance d'investigation.

Le logiciel étant construit sur une architecture centralisée, il peut être repensé sous la forme d'une architecture distribuée permettant la répartition de la charge, des données et des services.

Aspects applicatifs

Les décideurs doivent être capables de détecter rapidement les bons signaux porteurs d'informations utiles dans un contexte de stratégie d'anticipation. Nous pensons que la solution

proposée peut s'avérer pertinente pour investiguer dans des grandes masses de données, en particulier dans le contexte des lanceurs d'alerte.

Les tests réalisés ont pour objectif de valider les concepts et étudier la pertinence des solutions sur des jeux de données réels. Ils sont mis en œuvre notamment sur :

— une base de données bibliographiques construite sur l'état de l'art effectué par Mühlroth afin d'extraire éventuellement de nouveaux domaines non mis en avant par cet expert des *signaux faibles* et *tendances émergentes*.
— une base de documents scientifiques composée des résumés de projets H2020 et sur laquelle nous cherchons à déterminer les thèmes scientifiques qui sont actuellement à l'étude au niveau de la recherche européenne.

Les résultats obtenus démontrent qu'il est préférable (même si non rédhibitoire) d'utiliser des documents dans leur entièreté plutôt que des résumés de ceux-ci. Des bases de documents conséquentes et composées de documents de plusieurs octets fournissent un cadre de recherche idéale pour la chaine de traitement. Dans nos expérimentations, nous testons les limites de notre solution sur des bases documentaires contenant peu de documents de taille réduite (1 ou 2 ko).

Projets H2020

Ce corpus est constitué des résumés des projets H2020 sur la période 2014-2021 provenant de l'Open Data du service d'information sur la recherche et le développement (CORDIS) de l'Union européenne (dump du 06/05/2020). Le corpus initial, utilisé lors de la première étape de la chaine de traitement, permet d'identifier les thèmes *signaux faibles* potentiels sur les années 2014 et 2015. Nous étudions l'évolution des thèmes et des mots-clés associés sur les années suivantes.

Deux thèmes cohérents se sont démarqués : ils relèvent de la protection environnementale et des nouveaux matériaux. Deux thèmes supplémentaires ont émergé lors de l'itération suivante ayant trait au secteur médical, pour l'un, et à la biologie cellulaire, pour l'autre.

Perspectives. Cette expérimentation est réalisée dans une situation plutôt favorable car la base documentaire contient un grand nombre de documents sur des thèmes variés. Les documents sont cependant de taille réduite (1 ou 2 ko).
Il serait intéressant de tester l'évolution des thèmes détectés sur la base actualisée car des projets étaient toujours en cours de dépôt à la date du dump.

Analyse bibliographique des articles étudiés par Mühlroth

Dans cette base documentaire, nous utilisons les résumés d'articles présents dans l'état de l'art étudié par Mühlroth comme corpus initial de notre chaine de traitement. Le suivi

temporel est ensuite effectué sur les résumés d'articles sur la période d'avril 2017 à septembre 2020 obtenus avec une requête identique à celle de Mühlroth.

Le test décrit dans le chapitre 1 montre la difficulté de détecter des thèmes porteurs de *signaux faibles* potentiels sur une base documentaire de taille réduite (seulement 91 documents dans le corpus initial). Notre solution permet néanmoins la détection des prémisses d'un thème *signal faible* potentiel (celui-ci est confirmé comme tel sur la période d'analyse suivante). Sur la base complète, deux *signaux faibles* sont détectés portant sur deux domaines différents, mais proches : le premier est relatif à l'environnement, ses aspects météorologiques et climat, le second sur les matériaux et les impacts environnementaux engendrés.

Pour résumer, les prémisses d'un *signal faible* sont détectés sur le corpus initial (mots-clés associés au domaine médical). Ces prémisses sont confirmés à l'aide de documents de la seconde période : un *signal faible* potentiel relatif à ce domaine est détecté. Pour le confirmer comme *signal faible*, il serait nécessaire d'inclure une période supplémentaire de documents.
Sur la base documentaire associée aux deux périodes, deux *signaux faibles* potentiels sont donc détectés. Les deux relèvent de problématiques environnementales : météorologie, climat, nouveaux matériaux ; il s'agit de problématiques de recherche devenues essentielles actuellement.

Dans le second test décrit au chapitre 4, la solution proposée détecte sur le corpus initial (de taille réduite) un *signal faible* lorsque 14% de documents relatifs à ce thème sont présents. Ce seuil relativement élevé est dû à la nature des documents utilisés : il s'agit de résumés courts utilisant une terminologie plutôt générale.

L'utilisation de documents dans leur entièreté (et donc avec un vocabulaire associé plus pertinent et complet) devrait permettre un seuil de détection plus faible.

Perspectives. Contrairement à la base de documents issus de H2020, l'intérêt de la base documentaire bibliographique est qu'elle permet de tester les limites de notre approche. Il s'agit en effet de résumés de documents utilisant un vocabulaire plus générique que spécifique conforme à ce que l'on attend généralement d'un résumé écrit. La base est également de taille réduite : on ne peut donc pas parler de masses documentaires. Malgré ces conditions, le logiciel permet cependant de relever des signes précurseurs d'un *signal faible* (prémisses : sous la forme de mots-clés saillants), d'un *signal faible* potentiel (thème et mots-clés associés) et de *signaux faibles* avérés (grâce à un suivi longitudinal : un *signal faible* est considéré comme tel lorsqu'il devient *signal fort* dans le temps).
Pour valider notre approche, il est cependant nécessaire de consolider les résultats, notamment en nous appuyant sur des bases de documents conséquentes, indexés sur des périodes longues, et dont on peut aisément maîtriser la vérité terrain.

Annexe A

Critères et paramètrage des expérimentations

Pour les deux expérimentations sur les corpus de documents que sont les projets H2020 et la base d'articles étudiés par Mühlroth, nous détaillons l'ensemble des valeurs des paramètres utilisés dans le logiciel WILD.

TABLE *A.1* – ***Détail des paramètres de chaque composant et service utilisés dans la chaine de traitement du logiciel WILD. Pour chacun d'eux, nous donnons une description de leur objectif, des paramètres ajustables par l'utilisateur, leur importance et la phase de la chaine de traitement dans lesquelles ils interviennent.***

Service / Composant	**Paramètres**	**Projets H2020**	**Base d'articles Mühlroth**
Logiciel WILD	1. Nombre de fils d'exécution (threads) [instance]	1. 8	1. 8
LDA	1. Valeurs de *LDA* choisies	1. {2,3,4,5,6,7,8}	1. {2,3,4,5,6,7,8,9,10,11,12}
Word2Vec	1. Choix de la base d'apprentissage 2. Taille du vecteur caractéristique	1. Wikipédia 2. 300	1. Wikipédia 2. 300
w2vSim	1. Nombre de mots utilisés	1. 20	1. 40
Distance de Bhattacharyya	1. Seuil de conservation des liens entre thèmes	1. 0.6	1. 0.3
Modélisation thématique multi-niveaux			
tf-idf	1. Choix du calcul en fonction des caractéristiques recherchées	1. Equation 2.5	1. Equation 2.5

Suite à la page suivante

TABLE A.1 – suite de la page précédente

Service / Composant	Paramètres	Projets H2020	Base d'articles Mühlroth
Composant SMA associé aux documents pour attraction/répulsion	1. Fréquence d'actualisation du système	1. 60 fois/sec	1. 60 fois/sec
Agent "document"	1. Coefficient d'inertie 2. Coefficient de force 3. Probabilité de séléctionner un agent requête 4. Nombre d'itérations avant changement d'agent	1. 0.2 2. 0.7 3. 20% 4. 30	1. 0.2 2. 0.7 3. 20% 4. 30
Composant SMA associé aux mots pour attraction/répulsion	1. Fréquence d'actualisation du système	1. 60 fois/sec	1. 60 fois/sec
Agent "Mot"	1. Coefficient d'inertie 3. Coefficient de force 4. Probabilité de séléctionner un agent requête 5. Nombre d'itérations avant changement d'agent	1. 0.2 2. 0.7 3. 20% 4. 30	1. 0.2 2. 0.7 3. 20% 4. 30
Composant SMA pour la recherche Web	1. Fréquence d'actualisation du système	1. 60 fois/sec	1. 60 fois/sec
Agent de recherche	1. Nombre de mots voisins récupérés 2. Nombre de résultats de requêtes à récupérer	1. 2 2. tous	1. 2 2. tous
ProcessIterationService	1. Expression cron pour la périodicité de la tâche	1. "0 0 *? * * *"	1. "0 0 *? * * *"
SearchAgentService	1. Nombre d'agents par thème 2. Nombre de mots par requête 3. Nombre de requête avant oubli du meilleur agent	Désactivé	Désactivé
WebSearchService	1. Durée d'attente entre les requêtes Web 2. Liste des moteurs de recherche	1. 0 2. Elastic	1. 0 2. Elastic

Suite à la page suivante

TABLE A.1 – suite de la page précédente

Service / Composant	**Paramètres**	**Projets H2020**	**Base d'articles Mühlroth**
Composant de suivi historique des SMA associés aux mots	1. Nombre d'agents par thème 2. Nombre de mots récupérés par thème	1. 1 2. 1	1. 1 2. 1
SocketService			
SocketNewClientService	1. Numéro de port d'accès	1. 16000	1. 16000
DocumentRESTService	1. Numéro de port d'accès	1. 8081	1. 8081

Annexe B

Documents identifiés dans les domaines de la santé et du médical

Contenu des documents dont l'expérimentation dans le chapitre 1 a révélé des mots portant sur la thématique dans les domaines de la santé et du médical (cf. section 1.2.4.1.4). Ces documents sont présent dans le corpus initial des 91 articles référencés par Mühlroth.

In this paper, we detect emerging research fronts in a huge number of academic papers related to regenerative medicine, a field of radically innovative research. We divide citation networks into clusters using the topological clustering method, track the positions of papers in each cluster, and visualize citation networks with characteristic terms for each cluster. Analyzing the clustering results with the average published year and parent–child relationship of each cluster could be helpful in detecting recent trends. In addition, tracking topological measures, within-cluster degree z and participation coefficient P, enables us to determine whether there are emerging knowledge clusters. Our results show the success of our method in detecting emerging research fronts in regenerative medicine, and these results are confirmed as reasonable by experts. Finally, we predict the future core papers, with the potential of many citations, via the betweenness centralities in the citation network of the research into adult and somatic stem cells.

FIGURE *B.1* – ***Résumé de l'article de Shibata [SKT+11]***

Recently, health-related social media services, especially online health communities, have rapidly emerged. Patients with various health conditions participate in online health communities to share their experiences and exchange healthcare knowledge. Exploring hot topics in online health communities helps us better understand patients' needs and interest in health-related knowledge. However, the statistical topic analysis employed in previous studies is becoming impractical for processing the rapidly increasing amount of online data. Automatic topic detection based on document clustering is an alternative approach for extracting health-related hot topics in online communities. In addition to the keyword-based features used in traditional text clustering, we integrate medical domain-specific features to represent the messages posted in online health communities. Three disease discussion boards, including boards devoted to lung cancer, breast cancer and diabetes, from an online health community are used to test the effectiveness of topic detection. Experiment results demonstrate that health-related hot topics primarily include symptoms, examinations, drugs, procedures and complications. Further analysis reveals that there also exist some significant differences among the hot topics discussed on different types of disease discussion boards.

FIGURE *B.2* – ***Résumé de l'article de Lu [LZL+13]***

We proposed in this study to use anomaly detection models to discover research trends. The application was illustrated by applying a rule-based anomaly detector (WSARE), which was typically used for biosurveillance purpose, in the research trend analysis in social computing research. Based on articles collected from SCI-EXPANDED and CPCI-S databases during 2000 to 2013, we found that the number of social computing studies went up significantly in the past decade, with computer science and engineering among the top important subjects. Followed by China, USA was the largest contributor for studies in this field. According to anomaly detected by the WSARE, social computing research gradually shifted from its traditional fields such as computer science and engineering, to the fields of medical and health, and communication, etc. There was an emerging of various new subjects in recent years, including sentimental analysis, crowdsourcing and e-health. We applied an interdisciplinary network evolution analysis to track changes in interdisciplinary collaboration, and found that most subject categories closely collaborate with subjects of computer science and engineering. Our study revealed that, anomaly detection models had high potentials in mining hidden research trends and may provided useful tools in the study of forecasting in other fields.

FIGURE *B.3* – ***Résumé de l'article de Cheng [CLLH15]***

Traditional public health surveillance requires regular clinical reports and considerable effort by health professionals to analyze data. Therefore, a low cost alternative is of great practical use. As a platform used by over 500 million users worldwide to publish their ideas about many topics, including health conditions, Twitter provides researchers the freshest source of public health conditions on a global scale. We propose a framework for tracking public health condition trends via Twitter. The basic idea is to use frequent term sets from highly purified health-related tweets as queries into a Wikipedia article index - treating the retrieval of medically-related articles as an indicator of a health-related condition. By observing fluctuations in frequent term sets and in turn medically-related articles over a series of time slices of tweets, we detect shifts in public health conditions and concerns over time. Compared to existing approaches, our framework provides a general a priori identification of emerging public health conditions rather than a specific illness (e.g., influenza) as is commonly done.

FIGURE *B.4* – ***Résumé de l'article de Parker [PWY+13]***

The convergence of industries exposes the involved firms to various challenges. In such a setting, a firm's response time becomes key to its future success. Hence, different approaches to anticipating convergence have been developed in the recent past. So far, especially IPC co-classification patent analyses have been successfully applied in different industry settings to anticipate convergence on a broader industry/technology level. Here, the aim is to develop a concept to anticipate convergence even in small samples, simultaneously providing more detailed information on its origin and direction. Design/methodology/approach – The authors assigned 326 US-patents on phytosterols to four different technological fields and measured the semantic similarity of the patents from the different technological fields. Finally, they compared these results to those of an IPC co-classification analysis of the same patent sample. Findings – An increasing semantic similarity of food and pharmaceutical patents and personal care and pharmaceutical patents over time could be regarded as an indicator of convergence. The IPC co-classification analyses proved to be unsuitable for finding evidence for convergence here. Originality/value – Semantic analyses provide the opportunity to analyze convergence processes in greater detail, even if only limited data are available. However, IPC co-classification analyses are still relevant in analyzing large amounts of data. The appropriateness of the semantic similarity approach requires verification, e.g. by applying it to other convergence settings.

FIGURE *B.5* – ***Résumé de l'article de Preschitschek [PNLM13]***

At present, industries within the health and life science sector are moving towards one another resulting in new industries such as the medical nutrition industry. Medical nutrition products are specific nutritional compositions for intervention in disease progression and symptom alleviation. Industry convergence, described as the blurring of boundaries between industries, plays a crucial role in the shaping of new markets and industries. Assuming that the medical nutrition industry has emerged from the convergence between the food and pharma industries, it is crucial to research how and which distinct industry domains have contributed to establish this relatively new industry. The first two stages of industry convergence (knowledge diffusion and consolidation) are measured by means of patent analysis. First, the extent of knowledge diffusion within the medical nutrition industry is graphed in a patent citation interrelations network. Subsequently the consolidation based on technological convergence is determined by means of patent co-classification. Furthermore, the medical nutrition core domain and technology interrelations are measured by means of a cross impact analysis. This study proves that the medical nutrition industry is a result of food and pharma convergence. It is therefore crucial for medical nutrition companies to effectively monitor technological developments within as well as across industry boundaries. This study further reveals that although the medical nutrition industry's core technology domain is food, technological development is mainly driven by pharmaceutical/pharmacological technologies Additionally, the results indicate that the industry has surpassed the knowledge diffusion stage of convergence, and is currently in the consolidation phase of industry convergence. Nevertheless, while the medical nutrition can be classified as an industry in an advanced phase of convergence, one cannot predict that the pharma and food industry segments will completely converge or whether the medical industry will become an individual successful industry.

FIGURE *B.6* – ***Résumé de l'article de Weenen [WRP+13]***

As a basic knowledge resource, patents play an important role in identifying technology development trends and opportunities, especially for emerging technologies. However patent mining is restricted and even incomplete, because of the obscure descriptions provided in patent text. In this paper, we conduct an empirical study to try out alternative methods with Derwent Innovation Index data. Our case study focuses on nano-enabled drug delivery (NEDD) which is a very active emerging biomedical technology, encompassing several distinct technology spaces. We explore different ways to enhance topical intelligence from patent compilations. We further analyze extracted topical terms to identify potential innovation pathways and technology opportunities in NEDD.

FIGURE *B.7* – ***Résumé de l'article de Ma [MP15]***

The present study analyzes scientific publications on mass gatherings, characterizing its development as an emerging research field. We identified publications on mass gatherings, analyzing the scientific production and carrying out a co-citation analysis. We identified the works of reference that have laid the intellectual foundation for the field as well as the main scientific disciplines and journals that have contributed to its development. We identified 278 documents that cited 7149 bibliographic references. The 2006–2010 period saw a dramatic increase in the number of works published. Papers on mass gatherings also appeared frequently in multidisciplinary journals of high visibility and impact. The co-citation analysis revealed the existence of five clusters or thematic nuclei in research of the area. One large cluster brings together different studies on the prevalence of infectious diseases associated with pilgrimages to Mecca, and another cluster focuses on planning and response for health services in the context of mass gatherings associated with sporting events. Different indicators help characterize the nature of this emerging field, in which we observe the absence of a stable research community, the recentness of the bibliographic citations, and a high concentration of publications on the topic, with no peripheral areas of investigation. The study of mass gatherings is an emerging area of research with a notably multidisciplinary nature. Given the relevance and incidence of mass gatherings in relation to population health, it is necessary to foster the conditions that favor the consolidation of the field as a topic of research.

FIGURE *B.8* – ***Résumé de l'article de Gonzalez-Alcaide [GALR16]***

Annexe C

Paradigmes organisationnels dans les systèmes multi-agents

Différents paradigmes organisationnels ont été développés et chacun établit des modes de relation et d'interaction entre les agents. Nous présentons ici les principaux [HL04].

— **Hiérarchie**. Les agents sont hérarchisés selon la structure d'un arbre où chaque noeud représente un agent avec des liens d'autorité sur les noeuds-fils et de subordination avec les noeuds-parents. Les noeuds situés en haut de l'arbre ont une vision globale ; ils décomposent la tâche à fournir aux noeuds-fils. Les interactions ont lieu principalement entre les entités connectées selon la structure de l'arbre.

— **Holarchie**. Très similaire à la hiérarchie, l'holarchie se démarque par l'absence de relation d'autorité entre un agent et son sous-groupe. Les agents du sous-groupe constituent ensemble "physiquement" un sur-agent comme par exemple un banc de poissons ou un univers. Ce dernier comprend des galaxies, elles-même composées de systèmes solaires. Chaque groupe possède un caractère dérivé mais est distinct des entités qui le compose. En même temps, chaque groupe contribue aux propriétés du groupe supérieur. Comme la hiérarchie, l'holarchie s'applique facilement aux objectifs décomposables en sous-tâches.

— **Coalition**. La coalition est la formation d'un groupe temporaire dans le but d'accomplir un objectif commun dont la plus-value est supérieure à l'accomplissement individuelle de l'objectif. Chaque agent maximise son objectif individuel en formant un groupe avec d'autres agents. La structure d'organisation est principalement plate. Parmi elle, un agent "dirigeant", peut avoir un rôle de représentant pour l'ensemble du groupe. Une coalition peut ainsi être traitée comme une seule entité atomique.

— **Equipe**. Les agents travaillent ensemble à la réalisation d'un objectif commun. La différence avec la coalition est que les agents travaillent à maximiser l'intérêt commun plutôt qu'individuel. Ils se coordonnent dans leurs actions en cohérence avec l'objectif afin de le soutenir. Les agents assument un ou plusieurs rôles au sein de l'équipe pour répondre aux sous-tâches requises pour accomplir l'objectif donné à l'équipe.

— **Congrégation**. Les congrégations, similaires aux coalitions et équipes, sont des groupes d'agents travaillant en commun. A la différence des deux autres organisations, les congrégations sont permanentes et ont généralement plusieurs objectifs à atteindre. Les agents peuvent entrer ou sortir de ses congrégations et appartenir à plusieurs d'entre elles en même temps. Les congrégations ont une longue durée de vie et sont composées d'agents ayant des caractéristiques similaires ou complémentaires pour sélectionner des collaborateurs potentiels à la réalisation des objectifs.
— **Société**. La société est formée d'un ensemble d'agents avec des objectifs différents qui interagissent, communiquent. Ils sont soumis à un ensemble de loi commune. Similaire à nos sociétés, cette organisation ressemble à une construction sociale à longue durée. Ce paradigme d'organisation ouvert dans lequel les agents peuvent entrer et sortir à volonté est persistant. Cet environnement formé par la société permet les rencontres et interactions entre les entités qui la composent.
— **Fédération**. La fédération est formée d'agents cédant leur autonomie à un agent "délégué" du groupe qui les représente auprès des délégués des autres groupes. Les agents dans un groupe interagissent uniquement avec leurs délégués. Cette organisation est similaire au système gouvernemental dans lequelle des provinces régionales cèdent une partie de leur autonomie au gouvernement central unique et conservent une partie d'autonomie locale.
— **Marché**. Cet organisation fonctionne sur le principe de l'achat/vente. Des agents acheteurs font des offres sur des ressources, biens ou services proposés à la vente, par des agents vendeurs. Cette organisation simule des systèmes de marchés avec des statégies de négociation. Des agents peuvent aussi proposer à la vente des articles au marché où un agent "commissaire-priseur" est chargé de traiter les offres pour sélectionner la meilleure.
— **Matrice**. Similaire à une organisation hérarchique, l'organisation matricielle se différencie dans sa hérarchie où un agent peut être influencé par plusieurs agents. Ainsi la réussite de l'objectif d'un agent a des répercussions sur de multiples entités. Cette organisation est proche du fonctionnement humain.
— **Combinaison**. Cette organisation est le mélange de plusieurs d'entre-elles afin de répondre aux besoins spécifiques des objectifs donnés par l'utilisateur au système. Il peut s'agir d'une fédération de coalition ou une hiérarchie d'équipe. Ces systèmes peuvent avoir des organisations différentes pour le contrôle, les flux de données et d'autres parties du système. Ces organisations peuvent se chevaucher et s'imbriquer pour s'organiser de différentes manières selon une vision micro ou macro du système.

Annexe D

Expérimentation complémentaire projets H2020

Dans ce test, le corpus initial est constitué de l'ensemble des 28'145 résumés de projets H2020. Nous effectuons seulement la première étape de la chaine de traitement pour étudier les domaines de recherches principaux et leurs mots-clés (cf tableau 4.2).

Le tableau 4.2 montre les résultats obtenus sur le corpus H2020, années 2014 à 2021, après élagage, 8 thèmes sont retenus. Les thèmes H2020 sont bien identifiés :

— thème 1. Sémantique autour de la santé et du médical avec des mots-clés sur le cancer et les traitements avec par exemple comme mots-clés saillants : "RPC" pour les troubles rénaux, "NTM" pour les mycobactéries non tuberculeuses, "E-DURA" le nom d'implants neuronaux de motricité nouvelles générations,... ;
— thème 2. Cette thématique regroupe différents mots-clés de domaines portant sur les réseaux, les plateformes et les données avec par exemple comme mots-clés saillants : "C-ALM AOHE" pour échangeur de chaleur air-huile à couche additive manufacturée, "GlobalDNA" qui est un projet visant à offrir différentes vues sur les événements qui se produisent dans le monde à partir de données multimodales,... ;
— thème 3. Cette thématique est relative au domaine de la biologie cellulaire avec par exemple comme mots-clés saillants : "TopoII" qui est un ADN topoisomérases de type 2, "xylan" (Xylane en français) est un composant principal des hémicelluloses (composantes du bois),... ;
— thème 4. Les mots présents dans ce thème sont associés aux technologies et aux coûts pour l'industrie et les sociétés avec par exemple comme mots-clés saillants : le projet "WATTsUP" d'avion électrique, une suite logiciel "POBUCA" d'outils intelligents pour la fidélisation de la clientèle, le projet "MENDER" pour la mise à jour à distance des logiciels/micrologiciels de dispositifs intégrés,... ;
— thème 5. Cette thématique est relative aux théories de la physique, aux méthodes, aux

nouvelles technologies issues de la recherche en physique et biologie avec par exemple comme mots-clés saillants : “TopoII” et “xylan” (présentés ci-dessus), le multiplicateur de “Bogomolov” qui est un invariant géométrique, la microscopie à force de sonde Kelvin “KPFM” qui est une variante sans contact de la microscopie à force atomique,... ;

— thème 6. Domaine relatif aux projets européens et politiques ainsi que à la culture et aux projets sociaux avec par exemple comme mots-clés saillants : “sentience” sur la science de la sensibilité animale dans un projet de recherche interdisciplinaire, le programme d'investisseurs institutionnels étrangers qualifiés “QFIIs”, l'étude du nombre de publications scientifiques sur les méduses “JF” (jellyfish) et de leurs impacts sur l'industrie de la pêche et du tourisme,... ;

— thème 7. Cette thématique regroupe différents mots-clés de domaines portants sur l'énergie, l'environnement, la production et les déchets avec un point de vue industriel avec par exemple comme mots-clés saillants : “WPP” pour les parcs éoliens, “walnut” pour les arbres feuillus plantés dans des programmes de financement par l'UE, l'azote liquide “LN2” utilisé pour refroidir des détecteurs à semi-conducteur,... ;

— thème 8. La sémantique abordée dans ce thème est associée à la recherche scientifique européenne, aux partenaires européens, internationaux, la recherche en réseau et l'innovation avec par exemple comme mots-clés saillants : le projet “GetReal” qui vise à incorporer des données cliniques dans le développement de médicaments, le projet “PIANO” de coopération stratégique dans le domaine de l'eau, le programme “spaceEU” de sensibilisation et d'éducation à l'espace pour les jeunes, le projet d'interaction cerveau-ordinateur neuronal “BNCI”,...

Les thèmes 3, 4, 5 et 6 ont des documents en commun dû à la proximité des domaines abordés. Les thèmes 6 et 8 ont un seul document en commun, comme les thèmes 1 et 3. Les thèmes 2 et 7 sont sur des domaines unitaires puisqu'ils n'ont aucun document en commun.

	thème 1	**thème 2**	**thème 3**	**thème 4**	**thème 5**	**thème 6**	**thème 7**	**thème 8**
thème 1	19	0	1	0	0	0	0	0
thème 2	0	19	0	0	0	0	0	0
thème 3	1	0	16	19	21	19	0	0
thème 4	0	0	19	19	19	19	0	0
thème 5	0	0	21	19	17	19	0	0
thème 6	0	0	19	19	19	18	0	1
thème 7	0	0	0	0	0	0	20	0
thème 7	0	0	0	0	0	1	0	27

TABLE *D.1* – ***Nombre de documents communs entre thèmes.***

TABLE *4.2* – ***Expérimentation effectuée sur le corpus de documents relatifs aux projets H2020 pour les années 2014 à 2021. Présentation des résultats obtenus après construction de l'arbre et élagage. Pour chaque thème, est indiquée la valeur de*** k ***du*** LDA ***où il est détecté, la cohérence*** Word2Vec***, le nombre de documents associés ainsi que les 20 premiers mots triés par*** tf-idf ***et par*** LDA***.***

Nom du thème	thème 1				thème 2			
Nombre de documents	20				19			
Cohérence du thème (I_1)	147				139			
LDA $k =$	8				8			
	TF-IDF		LDA		TF-IDF		LDA	
Premiers mots du thème	rpc	0.41118	clinical	0.01112	c-alm	0.54279	data	0.01742
	ntm	0.39996	patients	0.01088	aohe	0.54279	project	0.00840
	e-dura	0.39787	cancer	0.00940	globaldna	0.39404	systems	0.00798
	oncomastr	0.39573	treatment	0.00861	5g-smart	0.39404	design	0.00620
	dux4-igh	0.38540	health	0.00798	fec	0.39153	technologies	0.00496
	rhabdoid	0.38463	disease	0.00769	kdd	0.38463	applications	0.00468
	rfta	0.37368	drug	0.00537	r-wake	0.35328	based	0.00448
	ccc	0.36471	project	0.00531	payload-lifting	0.35328	develop	0.00439
	lipofuscin	0.35948	development	0.00530	cybeco	0.35026	tools	0.00402
	rtks	0.35503	patient	0.00510	hipersim	0.34927	technology	0.00391
	stimos	0.34150	medical	0.00503	exa2pro	0.34380	control	0.00387
	cb1r	0.33945	diseases	0.00456	kconnect	0.33156	performance	0.00381
	iamd	0.33541	care	0.00415	hi-omics	0.32697	network	0.00378
	erfe	0.33146	therapy	0.00412	spaceports	0.32353	development	0.00373
	epitaxos	0.32952	risk	0.00375	ments	0.32199	software	0.00373
	mgus	0.32856	develop	0.00370	gaspils	0.32016	models	0.00362
	heartguide	0.32654	therapeutic	0.00350	body-based	0.31835	platform	0.00336
	atxa	0.32571	based	0.00346	tabede	0.31427	security	0.00326
	nocturne	0.32016	cells	0.00331	flysec	0.30912	provide	0.00311
	vitd	0.31427	early	0.00313	copepods	0.30766	integrated	0.00298

Suite à la page suivante

TABLE 4.2 – suite de la page précédente

Nom du thème	thème 3				thème 4			
Nombre de documents	38				38			
Cohérence du thème (I_1)	134				125			
LDA $k =$	8				8			
	TF-IDF		LDA		TF-IDF		LDA	
Premiers mots du thème	topoii	0.48786	cell	0.00973	wattsup	0.52204	market	0.02355
	xylan	0.41817	cells	0.00931	pobuca	0.47909	business	0.00887
	tbx18	0.41454	mechanisms	0.00606	mender	0.46853	technology	0.00831
	kmyr	0.40685	molecular	0.00602	tom-e	0.43137	solution	0.00768
	null	0.40625	understanding	0.00471	tactotek	0.41281	's	0.00645
	atabc	0.40177	brain	0.00459	null	0.40625	product	0.00600
	m1a	0.38906	human	0.00457	enroute	0.39404	company	0.00550
	centrosome/cilium	0.38332	genetic	0.00430	ivalo	0.37945	project	0.00508
	myo6	0.36950	role	0.00419	blindshell	0.37945	phase	0.00492
	gamma-turcs	0.36106	proteins	0.00415	loowatt	0.37945	costs	0.00440
	maits	0.36053	biology	0.00404	vaporpv	0.37482	years	0.00438
	t9ss	0.35878	development	0.00392	cortime	0.35601	time	0.00429
	trib1	0.35680	project	0.00392	endive	0.35328	companies	0.00401
	rtks	0.35503	dna	0.00392	paptic	0.34783	feasibility	0.00394
	evi1	0.35218	function	0.00380	innspect	0.34534	plan	0.00385
	fbxo11	0.34795	protein	0.00374	oblow	0.34359	cost	0.00384
	rab27a	0.34569	gene	0.00346	revault	0.34150	global	0.00376
	isg15	0.34479	cellular	0.00339	digistone	0.34150	industry	0.00367
	moz	0.34359	study	0.00336	mirrorpv	0.34047	platform	0.00347
	hb-egf	0.34150	functional	0.00335	rebelrocket	0.33899	commercial	0.00338

Suite à la page suivante

TABLE 4.2 – suite de la page précédente

Nom du thème	thème 5				thème 6			
Nombre de documents	38				38			
Cohérence du thème (I_1)	125				124			
LDA $k =$	8				8			
	TF-IDF		LDA		TF-IDF		LDA	
Premiers mots du thème	topoii	0.48786	quantum	0.00752	sentience	0.44683	project	0.01231
	xylan	0.41817	project	0.00692	null	0.40625	social	0.00848
	null	0.40625	materials	0.00628	qfiis	0.40309	's	0.00578
	bogomolov	0.37482	properties	0.00573	jf	0.40116	study	0.00521
	kpfm	0.35126	systems	0.00480	mereological	0.32931	climate	0.00494
	bijels	0.34265	field	0.00412	postgrowth	0.32116	understanding	0.00460
	tlss	0.32524	applications	0.00389	indo-aryan	0.32016	data	0.00451
	spss	0.32217	physics	0.00387	n400	0.32016	studies	0.00415
	amo-dance	0.31046	optical	0.00360	kelp	0.31916	change	0.00411
	silyliumylidene	0.29179	theory	0.00341	caricature	0.31817	analysis	0.00409
	igmf	0.28905	high	0.00335	neg	0.31235	cultural	0.00364
	iicn	0.28459	develop	0.00322	pdcw	0.31218	global	0.00358
	sloshing	0.27941	fundamental	0.00321	n-fe2o3	0.30735	political	0.00351
	topspin	0.27596	study	0.00320	bsr	0.30491	approach	0.00319
	ncss	0.27524	understanding	0.00318	playmoss	0.29840	knowledge	0.00316
	metalloporphyrins	0.27412	devices	0.00314	nuragic	0.29272	history	0.00302
	pil	0.27320	light	0.00305	dickens	0.28860	human	0.00300
	mwmi	0.26872	based	0.00300	choral	0.28198	policy	0.00292
	llps	0.26785	techniques	0.00293	civiciti	0.28165	europe	0.00280
	lensless	0.26525	methods	0.00289	knickpoint	0.27878	european	0.00272

Suite à la page suivante

TABLE 4.2 – suite de la page précédente

Nom du thème	thème 7				thème 8			
Nombre de documents	20				28			
Cohérence du thème (I_1)	122				118			
LDA $k =$	8				8			
	TF-IDF		LDA		TF-IDF		LDA	
Premiers mots du thème	wpp	0.53049	energy	0.01760	getreal	0.39639	project	0.01233
	walnut	0.42825	production	0.00937	piano	0.38540	innovation	0.01212
	ln2	0.37482	project	0.00885	spaceeu	0.37945	european	0.01153
	umbrellas	0.37412	water	0.00806	bnci	0.37317	training	0.00764
	healex	0.36919	technology	0.00686	train@ed	0.37071	support	0.00697
	agitator	0.36792	high	0.00659	glopid-r	0.36272	eu	0.00644
	eco-pet	0.36354	process	0.00627	supeera	0.35328	europe	0.00640
	rvcr	0.36106	materials	0.00541	prosme	0.34661	development	0.00582
	carbfix	0.33924	food	0.00486	e1st	0.33541	management	0.00579
	ice-protection	0.33591	power	0.00460	sdin	0.33263	services	0.00569
	carbafin	0.33049	industrial	0.00455	esq	0.32837	knowledge	0.00565
	ua-es	0.32524	environmental	0.00421	speakngi.eu	0.32697	activities	0.00557
	sensop-ii	0.32467	industry	0.00417	ehri	0.32437	researchers	0.00542
	waam	0.32016	cost	0.00404	postgrowth	0.32116	smes	0.00488
	rol	0.32016	waste	0.00403	aginfra	0.31427	network	0.00459
	lancey	0.31835	efficiency	0.00400	eubi	0.31392	partners	0.00457
	ngr	0.31621	market	0.00372	eo4agri	0.31046	public	0.00441
	aqs	0.31305	products	0.00363	bioeast	0.31046	scientific	0.00430
	saffron	0.31140	based	0.00360	eu-ibisba	0.30838	international	0.00426
	rencat	0.31046	innovative	0.00360	31st	0.30751	's	0.00420

Publications

Articles de journaux

Internationaux

— Julien Maitre, Michel Ménard, Guillaume Chiron, Alain Bouju. Détection de signaux faibles dans des masses de données faiblement structurées. Recherche d'Information, Document et Web Sémantique, ISTE OpenScience, 2019, 3 (1), 10.21494/ISTE.OP.2020.0463

Articles de conférences & workshops

Internationaux

— Julien Maitre, Michel Ménard, Guillaume Chiron, Alain Bouju, Nicolas Sidère. A meaningful information extraction system for interactive analysis of documents. International Conference on Document Analysis and Recognition (ICDAR 2019), Sep 2019, Sydney, Australia. pp.92-99, 10.1109/ICDAR.2019.00024

— Julien Maitre, Michel Ménard, Alain Bouju, Guillaume Chiron. Recherche de signaux faibles dans un contexte d'investigation numérique. Conférence Internationale H2PTM - Hypertextes et Hypermédias. Produits, Outils et Méthodes, Oct 2019, Montbéliard, France. pp.200-215

Nationaux

— Julien Maitre. Détection et analyse des signaux faibles pour un service caché Lanceurs d'alerte. APVP 2019 – l'Atelier sur la Protection de la Vie Privée, Jul 2019, Saint-Valery-sur-Somme

— Julien Maitre, Michel Menard, Guillaume Chiron, Alain Bouju. Utilisation conjointe LDA et Word2Vec dans un contexte d'investigation numérique. Extraction et Gestion des Connaissances 2017, Jan 2017, Grenoble, France

Table des figures

1 Aperçu d'un exemple de plateforme d'investigation. Dans le contexte du data journalisme, 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. 7

2 Stratégie de détection des signaux faibles. Elle passe par l'analyse de l'information précoce apportée par un lanceur d'alerte et l'extraction des collections de mots associées aux Thèmes découverts. Ces informations permettent ensuite de mieux cibler la phase de data mining pour la détection des signaux faibles. Chaque rectangle de couleur représente une source d'information (ex : documents) exploitée. 8

3 Détail des différentes étapes de l'extraction de connaissances/fouilles de données dans notre chaine de traitement ainsi que des propriétés et résultats obtenus. . . 9

4 Les différentes étapes pour extraire les signaux faibles 17

1.1 Détail des différentes étapes de l'extraction de connaissances/fouilles de données de notre chaine de traitement sur les articles de WoS ayant trait aux signaux faibles. 34

1.2 Représentation des traitements utilisés sur les différentes sources de données. Les résumés des articles référencés par Mühlroth [MG18] correspondent au corpus. Nous appliquons différentes techniques de nettoyage pour rendre les données exploitables par l'approche de modélisation thématique utilisée. Nous appliquons LDA, nous lançons plusieurs expérimentations en faisant varier le nombre de thèmes. Pour notre approche de plongement de mots qu'est Word2Vec, nous utilisons un dump des articles provenant du Wikipédia Anglais pour entrainer un modèle au moyen de l'outil Word2vec-on-wikipedia. 39

1.3 Représentation graphique du nombre de documents obtenu après requêtes des mots dans un moteur de recherche 51

1.4 Représentation graphique du nombre de documents attribués pour chaque thème (cf. 1.6). L'itération 0 correspond au document de l'étape de collecte de données du processus d'extraction de connaissances. L'itération 1 rajoute les documents obtenus après les requêtes effectuées par mots-clés dans un moteur de recherche (ElasticSearch). Les taux de croissance sont respectivement 8.7 (thème 1), 7.4 (thème 2), 24.3 (thème 3), 12.5 (thème 4), 6.3 (thème 5), 6.2 (thème 6), 7.3. (thème 7) et 6.0 (thème 8). 52

1.5 Frise chronologique de la méta-analyse bibliographique sur les articles de Mühlroth. Sur la période 1997-2017, la chaine de traitement permet de détecter un domaine médical et santé. Ce *signal faible* potentiel est confirmé après la récupération de nouveaux documents lors de l'étape d'agent mining. Deux autres signaux faibles potentiels sont aussi détectés portant, pour le premier sur le domaine environnement/climatologie/météorologie et pour le second sur celui des matériaux et leurs impacts environnementaux. 56

1.6 La première partie de la chaine de traitement utilise une approche de Topic Modeling et de Word Embedding. L'algorithme que nous proposons dans le chapitre 2 permet d'extraire les thèmes les plus pertinents sur plusieurs niveaux de LDA au moyen d'un indicateur basé sur Word2Vec. Le filtrage permet de récupérer les mots pertinents dans chaque thème selon notre définition du signal faible. . 59

1.7 La seconde partie de la chaine de traitement met en œuvre une solution d'agent mining, une combinaison de méthodes de data mining et de système multi-agents. Les documents du corpus sont projetés dans un système multi-agents animés par des forces d'attraction/répulsion construites sur des similitudes sémantiques (pour les documents) ou de contexte (pour les mots). Des agents de recherche enrichissent le système de nouveaux documents par des requêtes sur des moteurs de recherches (tel que Qwant) à partir des mots-clés des thèmes identifiés. 60

2.1 Représentation graphique de LSA 64

2.2 Représentation graphique de pLSA 65

2.3 Représentation graphique de LDA 66

2.4 Représentation graphique du plongement de mots 67

2.5 Projection en 2 dimensions de la similarité contextuelles de mots. Une valeur de similarité proche de 1 indique des mots utilisés dans des contextes similaires, inversement, 0 pour des mots avec un faible lien contextuel. 68

2.6 Exemple de termes (pays et leur capitale) projetés en deux dimensions par Analyse en composantes principales (PCA) appliquée à leurs descripteurs Word2Vec. Pendant l'apprentissage, aucune information n'est fournie sur le concept de capital. Le modèle est entrainé sur un corpus de Google News. Inspiré du travail de Mikolov [MSC+13, MCCD13] 69

2.7 Etape de génération du corpus de test sur la base d'un jeu de données extrait de Wikipédia. Nous appliquons LDA sur l'ensemble des documents, tout en faisant varier le nombre de thèmes. 72

2.8 Etape de caractérisation contextuelle des thèmes sur la base d'un modèle Word2Vec pré-entrainé. 72

2.9 La proposition 1 montre la recherche du niveau de l'arborescence donnant les thèmes les plus cohérents au sens du critère de ressemblance. La proposition 2 montre l'élagage des thèmes reliés entre eux dans l'arborescence au moyen d'un critère de ressemblance. L'objectif est dans ce deuxième cas de rechercher les thèmes les plus cohérents sur l'ensemble de l'arborescence. 73

2.10 Filtrage des thèmes au moyen d'une évalutation de la pertinence au moyen de tf-idf. Nous identifions les thèmes signaux faibles potentiels. 73

2.11 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. 74

2.12 Mise en œuvre de LDA sur un corpus de documents. En sortie, nous obtenons un ensemble de thèmes dont le nombre est spécifié en paramètre d'entrée. 75

2.13 Représentation de l'application de l'algorithme 1 sur les différentes partitions LDA obtenues pour $k \in \{2 \dots 5\}$. Les partitions sont organisées par niveau : la partition obtenue pour $k = 2$ est au niveau le plus haut. Les niveaux sont représentés par des rectangles. Les valeurs pour chaque thème sont celles de l'indicateur I_1. Les cercles rouge et vert représentent les plus petites valeurs de chaque niveau de LDA. Parmi ces valeurs, la plus élevée correspond à LDA 4 ($k = 4$), désignant la partition la plus cohérente au sens du critère contextuel. . 77

2.14 Représentation de l'algorithme 2 sur les différentes partitions LDA obtenues pour $k \in \{2 \dots 5\}$ après avoir construit l'arborescence en utilisant l'indicateur de ressemblance I_2. Un lien est formé entre chaque thème de niveau k et un thème de niveau $k + 1$ si la valeur de l'indicateur I_2 est supérieure à un seuil défini arbitrairement. Les partitions sont organisées par niveau : la partition obtenue pour $k = 2$ est au niveau le plus haut. Les niveaux sont représentés par des rectangles. Les thèmes ont tous été triés dans une liste par ordre croissant selon l'indicateur I1 sans distinction du niveau k (cf. Tableau 2.2). Nous sélectionnons le premier élément de la liste et supprimons ses fils et parents avec les méthodes $Parents(c_k)$ et $Fils(c_k)$. Cette tâche est exécutée jusqu'à ce que la liste soit vide. Les éléments récupérés forment les thèmes les plus pertinents selon l'algorithme 2. 79

2.15 Échantillon d'un document. Les couleurs représentent les groupes de mots. Le document contient des mots du Thème principal 2 81

2.16 Construction des 4 corpus de tests artificiels. Les mots du Thème *signal faible* sont injectés en quantité variable dans chaque document du corpus (test 1), ou bien dans une quantité variable de documents (test 2), ou encore en quantité variable dans un seul document avec (test 3) ou sans mots-outils (test 4) 82

2.17 Expérimentation 1 : résultats de l'application de l'algorithme 2 comparativement aux 6 LDA seuls paramétrés avec des valeurs de K allant de 5 à 8. Sur chaque ligne est donné le nombre de mots, #w, relatif au signal faible inséré dans chaque document "SecretWords-#w". 83

2.18 Expérimentation 2 : résultats de l'application de l'algorithme 2 comparativement aux 6 LDA seuls paramétrés avec des valeurs de K allant de 5 à 8. Sur chaque ligne est décrit le nombre de documents, #d, porteurs du signal faible (i.e. contenant 2500 mots relatifs au Thème signal faible). 85

2.19 Expérimentation 2 Bis : résultats de l'application de l'algorithme 2 comparativement aux 6 LDA seuls paramétrés avec des valeurs de K allant de 5 à 8. Sur chaque ligne est décrit le nombre de documents, #d, porteurs du signal faible (i.e. contenant 600 mots relatifs au signal faible). 86

2.20 Expérimentation 3 : résultat de l'application de l'algorithme 2 comparativement aux 9 LDA seuls paramétrés avec des valeurs de K allant de 2 à 8. Sur chaque ligne est donné le nombre de mots, #w, appartenant au Thème relatif au signal faible dans le document "SecretWords-#w". 87

2.21 Expérimentation 4 : 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. 90

2.22 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 leur distribution originale. Ces mots non-communs sont identifiés par une étude de co-occurrence entre Thèmes. 92

2.23 Résultats de l'algorithme 2 comparé aux 7 LDAs originaux paramétrés avec k variant de 2 à 8 sur la détection d'un thème signal faible dans les thèmes déterminés par l'approche de modélisation thématique multi-niveaux. Dans chaque document, nous insérons 3 séries de 4 mots du Thème Droit (signal faible). L'identification du thème signal faible est réalisée par le calcul de la somme des poids des mots n'appartenant qu'à un des 5 Thèmes (Economie, Histoire, Informatique, Médecine et Droit (Signal faible)). Le thème est identifié signal faible lorsque la valeur du Thème Droit est la plus importante. 93

2.24 Moyenne sur 10 tests de la valeur de cohérence sémantique du thème signal faible "détecté" avec l'algorithme 2 comparée aux 7 LDAs originaux paramétrés avec k variant de 2 à 8. L'algorithme 2 détecte le thème signal faible avec la plus grande valeur de cohérence comparée à celles de LDA. 94

2.25 Un échantillon de document original provenant du corpus Ohsumed. 94

2.26 Un échantillon de document original provenant de Wikipedia et traitant de maladies connexes à Ebola. 95

2.27 Liste des mots du corpus ayant les plus grandes valeurs de pondération tf-idf. On remarque que les mots clés du thème signal faible détecté sont parmi les premiers de la liste. 97

3.1 Le processus d'extraction de connaissances 101

3.2 Carte heuristique des approches pour de la fouille du contenu Web. Nous détaillons les différents niveaux de structuration et formes de données du Web. On peut identifier deux groupes de méthodes parmi les techniques de fouille du contenu du Web : “Fouille du contenu de pages web” et “Fouille de résultats de recherche”. Les techniques principalement employées sont la classification, le regroupement ainsi que le résumé automatique de texte. Dans notre étude, nous nous positionnons selon les cadres rouges. Les agents de recherche extraient de l'information à partir de données non structurées. Celles-ci permettent de suivre des *signaux faibles* potentiels. A partir de résultats sur des moteurs de recherche, les agents de recherche fouillent le contenu de pages Web et créent de nouveaux agents. Ces agents se classifient dans un système multi-agents à partir de l'environnement de ce dernier comprenant les agents déjà présents. 106

3.3 Carte heuristique des approches de la fouille de la structure du Web. Ce schéma détaille les différentes applications, algorithmes, utilisations et représentations dans les approches d'extraction d'information sur la structure du web. Deux algorithmes largement employés existent : “PageRank” et “HITS”. Les approches, pour la fouille de la structure du Web, permettent un nombre important d'applications et sont principalement employées dans des moteurs de recherche ainsi que pour l'analyse de réseaux sociaux. La représentation couramment utilisée associe les entités à des noeuds et les flux ou relation à des liens. 108

3.4 Carte heuristique des approches pour la fouille de l'usage du Web. Dans la fouille de jounaux du web, trois phases principales sont utilisées. Le prétraitement, rassemblant un ensemble de techniques allant du nettoyage des données à l'identification des utilisateurs, transforme les logs d'utilisateurs en structure formatée pour de futur traitement avec des méthodes de data mining. La “découverte de modèles” recherche des modèles intéressants à appliquer sur des données de journaux de serveurs au moyen de différentes méthodes. La dernière phase, l'analyse de modèle, appliquée après les méthodes de data mining doit permettre de comprendre les résultats pour une prise de décision. 109

3.5 Schéma de la chaine de traitement. Les thèmes obtenus par la modélisation thématique multi-niveaux décrite au chapitre 2 ainsi que le corpus de documents sont injectés dans les systèmes multi-agents. Une interface visuelle et interactive permet à l'utilisateur d'interagir avec le système. Ce dernier effectue alors des recherches sur le Web afin d'enrichir le corpus de nouveaux documents en rapport avec les thèmes trouvés et ceux sélectionnés par l'utilisateur. Lorsque l'utilisateur choisi de manière interactive un document, les requêtes sont construites à partir des mots-clés présents dans le document et découverts dans les thèmes. Actuellement, l'interface utilisateur permet uniquement une intéraction classique simple de type clic souris. 117

3.6 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 aux différents thèmes obtenus lors de la modélisation thématique multi-niveaux et éventuellement présents dans les documents sélectionnés par l'utilisateur (renforcement) (cf. Chapitre 2). 118

3.7 Pour chaque document est calculé un vecteur caractéristique composé de n valeurs, chacune d'elles définit la composante du document associée à un thème. Ce vecteur caractéristique permet l'application des forces d'attraction/répulsion dans le système multi-agents construit. Pour chaque thème, et chaque document, nous calculons le nombre d'occurence des mots du thème présents dans le document en prenant en compte leur rareté et le fait que ces mots soient présents dans peu de documents (grâce à leur valeur de *tf-idf*). 119

3.8 Chaque agent “document” a un vecteur caractéristique qui décrit les liens du document avec chaque thème. Ces liens servent à calculer les forces d'attraction et répulsion de l'agent “document” A. La somme des forces résultantes s'ajoute à celles calculées aux itérations précédentes. 120

3.9 Représentation sous Unity de l'évolution des agents “document” dans un environnement 3D en fonction des forces d'attraction/répulsion qui les animent. La figure 3.11a représente les agents dans leur état initial. Grâce aux forces de similarité/dissimilarité entre agents, ceux-ci commencent par s'éloigner les uns des autres (cf. figures 3.11b et 3.9c) puis se regroupent entre agents “document” similaires, et commencent à former des clusters (cf. figure 3.9d), pour enfin tomber dans des états stables (cf. figures 3.9e et 3.9f). 122

3.10 L'utilisateur visualise les documents du système multi-agents dans un espace en 3D. Chaque document est représenté par un point dans l'espace. L'utilisateur peut sélectionner un agent pour le transformer en agent-requête. Les autres agents de l'espace vont s'organiser autour de cet agent que l'utilisateur aura sélectionné en utilisant les forces d'attraction/répulsion calculées entre paires d'agents à partir des N composantes des vecteurs caractéristiques. 123

3.11 Représentation sous Unity d'un agent-requête (couleur blanc). La figure 3.11a représente les agents dans leur état stable. Un agent de couleur bleu est transformé en agent-requête : il génère des forces d'attraction/répulsion qui attirent/repoussent les autres agents (cf. figure 3.11b). 123

3.12 Un agent de recherche effectue des requêtes dans un moteur de recherche avec des mots-clés d'un thème (e.g. signal faible potentiel). Pour chaque document trouvé, il extrait le segment de document le plus pertinent contenant les mots-clés de la requête et l'ajoute aux documents déjà présents dans le corpus d'étude afin d'enrichir les données. 124

3.13 L'agent de recherche identifie la position des mots-clés (en vert) dans le document. Il extrait du document le segment de texte le plus pertinent contenant l'ensemble des mots-clés. Dans cet exemple, la requête était composée des mots-clés "innovation", "trends" et "technical". La section de couleur rouge correspond à la section récupérée pour former un nouvel agent "document". 125

3.14 Pour chaque combinaison du mot-clé avec un ou plusieurs mots du segment extrait, une entrée dans la liste est créée et sa valeur d'incrément augmentée suivant les résultats donnés par les recherches successives effectuées par chaque agent. 125

3.15 Vue globale du logiciel WILD et des différents services connectés servant d'interface pour l'accès au Web. 129

3.16 Vue détaillée de toutes les instances d'investigation, composants et services du logiciel WILD. Sont présentées également, les interfaces d'accès et de recherche : Interfaces Unity et Web et service d'accès aux moteurs de recherche. 130

3.17 Présentation des échanges de données entre les agents de recherche, le service "WebSearchService", le service TOR et les moteurs de recherche. Les agents effectuent leurs requêtes auprès du service de recherche Web. Celui-ci lance la requête sur les moteurs de recherche au travers du service TOR. 132

3.18 Un agent de recherche utilise les liens Web fournis par le service de recherche Web, suite à ses requêtes, pour tranformer en agents "document" les pages Web après extraction du contenu pertinent (voir 3.3.3.1) 133

3.19 Présentation des échanges de données via une socket entre un client de visualisation sous Unity et le logiciel WILD. Le service écoute sur un port, attendant la connexion de clients et le choix des informations sur l'instance qu'il souhaite afficher. La connexion est passée au service qui échange ces informations avec le client connecté. 134

3.20 Le service du processus itératif agit comme un orchestrateur et planifie la phase de modélisation thématique. Les thèmes sont mis à jour et les mots associés à ces thèmes sont de nouveau injectés dans le composant multi-agent associé aux mots. 137

3.21 Le service prend en entrée un ensemble de documents contenant un ou plusieurs signaux faibles potentiels. Les thèmes sont mis à jour et les listes de mots associés sont de nouveaux injectées dans le composant multi-agent associé aux mots. . . 138

3.22 Le service d'agents de recherche produit, à partir de groupes de mots provenant des thèmes, de nouveaux agents de recherche. Ces agents sont en charge de la récupération des sections de texte pertinentes afin d'augmenter le corpus initial. Ce corpus est ensuite ré-analysé par le service du processus itératif (cf. section 3.4.2.4) qui met en œuvre l'approche de modélisation thématique multi-niveaux (approche conjointe *LDA/Word2Vec*). Les thèmes et leurs mots-clés associés sont alors mis à jour et de nouvelles recherches sont lancées. 139

3.23 Le service prend en entrée chaque thème et leur liste de mots associées, et produit des combinaisons de groupe de mots afin de créer de nouveaux agents de recherche. Ces agents transmettent les groupes de mots sous la forme de requête au service de recherche Web (cf. section 3.4.1.3) pour récupérer les résultats de recherche. Les sections de texte résultantes des pages Web apparaissant dans les résultats de recherche définissent alors de nouveaux agents "document". 140

3.24 Le composant effectuant le suivi historique sauvegarde les thèmes et leur liste de mots-clés afin d'étudier leur évolution. Ces mots-clés permettent la construction de nouveaux agents de recherche. 140

3.25 Le composant effectuant le suivi prend en entrée les thèmes et les listes de mots-clés apportés par le composant SMA associé aux mots. Pour chacun d'entre eux, des agents de recherche sont créés à partir de combinaison de groupes de mots. Le composant enregistre l'évolution des valeurs tf-idf dans les logs. 141

3.26 Présentation fonctionnelle de la chaine de traitement et du sens de transfert des données entre fonctions. 142

4.1 Le site d'information CORDIS regroupe les news portant sur les projets H2020 en 11 grands domaines (Thèmes) (en date du 09/07/2020). 150

4.2 Evolution du nombre de documents rapportés par les mots-clés de chaque thème. Le corpus initial regroupe les articles des années 2014-2015. Nous indiquons le nombre de documents rapportés qui contiennent au moins un mot parmi les 20 premiers mots-clés, et ceci sur chacun des index. 157

4.3 Evolution de la valeur de tf-idf pour chaque thème. Le corpus initial regroupe les articles des années 2014-2015. La valeur est calculée sur les 20 premiers mots de chaque thème après qu'ils aient été triés. 158

4.4 Représentation graphique du nombre de documents obtenus après requêtes des mots dans un moteur de recherche. Seuls les mots rapportant un nombre de documents significatifs sont présents. 159

4.5 Evolution, pour chaque index, du nombre de documents rapportés par les mots-clés les plus pertinents. 162

4.6 Composants et services utilisés pour l'expérimentation sur la base documentaire de Mühlroth 166

B.1 Résumé de l'article de Shibata [SKT+11] 179

B.2 Résumé de l'article de Lu [LZL+13] 180

B.3 Résumé de l'article de Cheng [CLLH15] 180

B.4 Résumé de l'article de Parker [PWY+13] 181

B.5 Résumé de l'article de Preschitschek [PNLM13] 181

B.6 Résumé de l'article de Weenen [WRP+13] 182

B.7 Résumé de l'article de Ma [MP15] 182

B.8 Résumé de l'article de Gonzalez-Alcaide [GALR16] 183

Liste des tableaux

1.1 Chronologie des événements marquants dans la définition des signaux faibles . . 25

1.2 Chronologie des événements marquants pour les tendances 27

1.3 Requête de recherche, mots-clés utilisés, nombre de résultats obtenus et période sur laquelle s'applique la recherche 37

1.4 Détail des thèmes obtenus par notre approche de modélisation thématique multi-niveaux. Le thème 1 semble le plus cohérent d'après l'indicateur construit via l'approche Word2Vec. 45

1.5 Nombre de documents obtenus après requêtes à partir de mots-clés dans un moteur de recherche (ElasticSearch) composé des 537 résumés d'articles. Ces derniers couvrent la période 2017-2020 et ont été extraits à partir de la même requête utilisée par Mühlroth [MG18] sur la base de données de WoS (cf. Tableau 1.3). Cette liste comprend les requêtes qui ont permis la récupération d'au moins 15 documents supplémentaires. 49

1.6 Nombre de documents attribués à chaque thème. L'itération 0 correspond au document de l'étape de collecte de données du processus d'extraction de connaissances. L'itération 1 rajoute les documents obtenus après les requêtes effectuées par mots-clés dans un moteur de recherche (ElasticSearch). 50

1.7 Liste des thèmes et des mots-clés associés obtenus par modélisation thématique multi-niveaux combinant l'application conjointe d'une modélisation thématique et d'un plongement lexical guidée par une mesure de cohérence. Les mots de couleur verte ont permis la récupération d'au moins 15 documents supplémentaires. 53

1.8 Détail des thèmes 1, 3 et 4 obtenus par notre approche de modélisation thématique multi-niveaux après une nouvelle itération sur le nouveau corpus de documents. Le thème 1 porte sur le domaine de l'environnement climatique et météorologique, le thème 3 sur les matériaux et leurs impacts environnementaux et le thème 4 sur le domaine de la santé 54

2.1 Liste des 5 premiers mots de chaque thème 75

2.2 Liste des thèmes triés par valeur de l'indicateur I_1 selon l'exemple en figure 2.14. La sélection des thèmes résultats est faite par suppression des éléments de la liste selon les relations entre les thèmes décrits par l'indicateur I_2. Tant que la liste est non vide, le premier élément est sélectionné comme thème résultat et ses fils et parents (avec les méthodes $Parents(c_k)$ et $Fils(c_k)$) sont supprimés de la liste selon l'arborescence 80

2.3 Description du jeu de données. Dans le corpus initial, chaque document est composé intégralement de mots appartenant au champ lexical d'un même Thème. . 81

2.4 Composition du corpus généré pour le test (en mots) 88

2.5 Expérimentation sur un corpus de comptes-rendus médicaux avec ajout d'un signal faible sous forme de documents portant sur des agents infectieux. Présentation des résultats obtenus sur les 5 premiers thèmes (parmi les 13 thèmes triés selon leurs cohérences) obtenus après construction de l'arbre et élagage. Pour chaque thème, est indiquée 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. . . 96

3.1 Description des caractéristiques du Web [LC04, DC12, HKD13] 105

3.2 Tableau des catégories d'agent selon leur comportement et leur type. 113

3.3 Tableau des différents systèmes multi-agents appliqués aux documents, aux mots, à la recherche et description des différents agents présents dans les 3 systèmes. . 114

3.4 Détail des paramètres de chaque composant et service utilisés dans la chaine de traitement du logiciel WILD. Pour chacun d'eux, nous donnons une description de leur objectif, des paramètres ajustables par l'utilisateur, leur importance et la phase de la chaine de traitement dans lesquelles ils interviennent. 144

4.1 Expérimentation effectuée sur le corpus de documents relatifs aux projets H2020 pour les années 2014 et 2015. Présentation des résultats obtenus après construction de l'arbre et élagage. Pour chaque thème, est indiquée la valeur de k du LDA où il est détecté, la cohérence Word2Vec, le nombre de documents associés ainsi que les 20 premiers mots triés par tf-idf et par LDA. 152

4.2 Nombre de documents communs entre thèmes. 156

4.3 Tableau montrant le nombre de documents rapportés par les mots-clés de chaque thème. Une colonne détaille le rapport du nombre de documents rapportés à partir du dernier index sur le nombre de documents du corpus initial pour chaque thème. 158

4.4 Nombre de documents obtenus après requêtes à partir de mots-clés dans un moteur de recherche (ElasticSearch) composé des 23 060 résumés des projets H2020. Ces derniers couvrent la période 2016-2020 et ont été extraits à partir de l'Open Data du service CORDIS de l'union européenne (cf. section 4.2.1) 160

4.5 Présentation de deux thèmes émergents du thème 2 présent dans le tableau 4.1 obtenus après construction de l'arbre et élagage lors de l'itération suivante. Le corpus contient les documents initiaux ainsi que ceux récupérés dans l'index h2020-2016. Pour chaque thème, est indiquée la valeur de k du LDA où il est détecté, la cohérence Word2Vec ainsi que les 20 premiers mots triés par tf-idf et par LDA. 163

4.6 Nombre de documents communs entre thèmes après analyse sur les 8 années. . . 163

4.7 Constitution de la base de documents utilisée à l'initialisation de la chaine de traitement (base initale) et ajout d'un index ElasticSearch pour les requêtes Web. 164

4.8 Répartition du nombre d'articles utilisés dans les travaux de Mühlroth selon les 4 domaines de la classification PEST. 165

4.9 Détail du thème obtenu par notre approche de modélisation thématique multi-niveaux sur le corpus initial contenant 6 documents supplémentaires portant sur le domaine médical et santé. 168

A.1 Détail des paramètres de chaque composant et service utilisés dans la chaine de traitement du logiciel WILD. Pour chacun d'eux, nous donnons une description de leur objectif, des paramètres ajustables par l'utilisateur, leur importance et la phase de la chaine de traitement dans lesquelles ils interviennent. 175

D.1 Nombre de documents communs entre thèmes. 188

4.2 Expérimentation effectuée sur le corpus de documents relatifs aux projets H2020 pour les années 2014 à 2021. Présentation des résultats obtenus après construction de l'arbre et élagage. Pour chaque thème, est indiquée la valeur de k du LDA où il est détecté, la cohérence Word2Vec, le nombre de documents associés ainsi que les 20 premiers mots triés par tf-idf et par LDA. 189

Liste des Algorithmes

1 Recherche du niveau de l'arborescence donnant les thèmes les plus cohérents . . . 77

2 Récupération des thèmes pertinents dans l'arborescence LDA 78

Bibliographie

[AA15] Rubayyi Alghamdi and Khalid Alfalqi. A Survey of Topic Modeling in Text Mining. *IJACSA) International Journal of Advanced Computer Science and Applications*, 6(1) :147–153, 2015.

[AIS93] Rakesh Agrawal, Tomasz Imieliński, and Arun Swami. Mining Association Rules Between Sets of Items in Large Databases. *ACM SIGMOD Record*, 22(2), 1993.

[Ans75] H Igor Ansoff. Managing Strategic Surprise by Response to Weak Signals. *California Management Review*, 18(2) :21–33, 1975.

[Ans80] H. Igor Ansoff. Strategic issue management. *Strategic Management Journal*, 1980.

[Ans85] H. Igor Ansoff. Conceptual underpinnings of systematic strategic management. *European Journal of Operational Research*, 1985.

[AOGS13] Mariam Adedoyin-Olowe, Mohamed Medhat Gaber, and Frederic Stahl. TRCM : A methodology for temporal analysis of evolving concepts in Twitter. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 7895 LNAI, 2013.

[AP98] James Allan and Ron Papka. On-line New Event Detection. *SIGIR '98 Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 37–45, 1998.

[APLB05] Julien Ah-Pine, Julien Lemoine, and Hamid Benhadda. 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, 2005.

[APM+13] Luca Maria Aiello, Georgios Petkos, Carlos Martin, David Corney, Symeon Papadopoulos, Ryan Skraba, Ayse Goker, Ioannis Kompatsiaris, and Alejandro Jaimes. Sensing trending topics in twitter. *IEEE Transactions on Multimedia*, 15(6), 2013.

[AT10] Hidenao Abe and Shusaku Tsumoto. Trend detection from large text data. In *Conference Proceedings - IEEE International Conference on Systems, Man and Cybernetics*, 2010.

[AZK14] Assad Abbas, Limin Zhang, and Samee U. Khan. A literature review on the state-of-the-art in patent analysis, 2014.

[BAB13] Ahmad Barirani, Bruno Agard, and Catherine Beaudry. Discovering and assessing fields of expertise in nanomedicine : A patent co-citation network perspective. *Scientometrics*, 94(3), 2013.

[Bak18] Amir Bakarov. A Survey of Word Embeddings Evaluation Methods. 2018.

[BCE16] Tal Baumel, Raphael Cohen, and Michael Elhadad. Sentence Embedding Evaluation Using Pyramid Annotation. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 145–149. 2016.

[BCP+08] A Bergholz, JH Chang, G Paass, F Reichartz, and S Strobel. Improved Phishing Detection using Model-Based Features. *Ceas*, 2008.

[BCR15] Ramakrishna B. Bairi, Mark Carman, and Ganesh Ramakrishnan. On the Evolution of Wikipedia : Dynamics of Categories and Articles. *2015 ICWSM Workshop*, pages 1–8, 2015.

[BEG09] Levent Bolelli, Şeyda Ertekin, and C. Lee Giles. Topic and trend detection in text collections using latent dirichlet allocation. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 5478 LNCS, 2009.

[Ber00] Henry Russell Bernard. *Social research methods : Qualitative and quantitative approaches*. SAGE, 2000.

[BG18] Amir Bakarov and Olga Gureenkova. 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, pages 16–21. Springer Verlag, 2018.

[BGJT04] David M. Blei, Thomas L. Griffiths, Michael I. Jordan, and Joshua B. Tenenbaum. Hierarchical topic models and the nested Chinese restaurant process. In *Advances in Neural Information Processing Systems*. Neural information processing systems foundation, 2004.

[BGL14] Mohit Bansal, Kevin Gimpel, and Karen Livescu. 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, pages 809–815, 2014.

[Bha43] A Bhattacharyya. On A Measure of Divergence Between Two Statistical Populations Defined by their Probability Distributions. *Bulletin of the Calcutta Methematical Society*, 35(1) :99–109, 1943.

[BI06] Khoo Khyou Bun and Mitsuru Ishizuka. Emerging topic tracking system in WWW. *Knowledge-Based Systems*, 19(3), 2006.

[BL06] David M. Blei and John D. Lafferty. Dynamic topic models. In *ACM International Conference Proceeding Series*, volume 148, pages 113–120, 2006.

[BNJ03] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3(4-5) :993–1022, may 2003.

[BOMOC14] Gema Bello-Orgaz, Héctor Menéndez, Shintaro Okazaki, and David Camacho. Combining social-based data mining techniques to extract collective trends from twitter. *Malaysian Journal of Computer Science*, 27(2), 2014.

[BPK+01] Glenn David Blank, William M. Pottenger, G. Drew Kessler, Martin Herr, Harriet Jaffe, and Soma Roy. CIMEL (poster session) : constructive, collaborative inquiry-based multimedia E-learning. (January) :179, 2001.

[Bra16] Max Bramer. *Introduction to Data Mining*. 2016.

[BXMH15] Bing Kun Bao, Changsheng Xu, Weiqing Min, and Mohammod Shamim Hossain. Cross-platform emerging topic detection and elaboration from multimedia streams. *ACM Transactions on Multimedia Computing, Communications and Applications*, 11(4), 2015.

[BXZ+09] Shenghua Bao, Shengliang Xu, Li Zhang, Rong Yan, Zhong Su, Dingyi Han, and Yong Yu. Joint emotion-topic modeling for social affective text mining. In *Proceedings - IEEE International Conference on Data Mining, ICDM*, pages 699–704, 2009.

[Cav16] Federico Caviggioli. Technology fusion : Identification and analysis of the drivers of technology convergence using patent data. *Technovation*, 55-56, 2016.

[CCS13] Mario Cataldi, Luigi Di Caro, and Claudio Schifanella. Personalized emerging topic detection based on a term aging model. *ACM Transactions on Intelligent Systems and Technology*, 5(1), 2013.

[CGM09] Longbing Cao, Vladimir Gorodetsky, and Pericles A. Mitkas. Agent mining : The synergy of agents and data mining, 2009.

[Che06] Chaomei Chen. CiteSpace II : Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for Information Science and Technology*, 57(3), 2006.

[CL11] Clive Steven Curran and Jens Leker. Patent indicators for monitoring convergence - examples from NFF and ICT. *Technological Forecasting and Social Change*, 78(2), 2011.

[CLLH15] Qing Cheng, Xin Lu, Zhong Liu, and Jincai Huang. Mining research trends with anomaly detection models : the case of social computing research. *Scientometrics*, 103(2), 2015.

[CN09] Lawrence P Carr and Alfred J Nanni Jr. *Delivering results : managing what matters*. Springer Science Business Media, 2009.

[Cof97a] Brian Coffman. Weak signal research, part I : Introduction. *Journal of Transition Management*, 1997.

[Cof97b] Brian Coffman. Weak signal research, part II : Information theory. *Journal of Transition Management*, 1997.

[Cof97c] Brian Coffman. Weak signal research, part III : Sampling, uncertainty and phase shifts in weak signal evolution. *Journal of Transition Management*, 1997.

[Cof97d] Brian Coffman. Weak signal research, part IV : Evolution and growth of the weak signal to maturity. *MG Taylor. Available : http ://www. mgtaylor. com/mgtaylor/jotm/winter97/wsrmatur. htm*, 1997.

[Cof97e] Brian Coffman. Weak signal research, part V : A process model for weak signal research. *Journal of Transition Management*, 1997.

[CTT06] Yun Chi, Belle L. Tseng, and Junichi Tatemura. Eigen-trend : Trend analysis in the blogosphere based on singular value decompositions. In *International Conference on Information and Knowledge Management, Proceedings*, 2006.

[CWB+11] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12 :2493–2537, aug 2011.

[CWL10] Pao Long Chang, Chao Chan Wu, and Hoang Jyh Leu. Using patent analyses to monitor the technological trends in an emerging field of technology : A case of carbon nanotube field emission display. *Scientometrics*, 82(1), 2010.

[CWY12] Longbing Cao, Gerhard Weiss, and Philip S. Yu. A brief introduction to agent mining, 2012.

[CYK+11] Sungchul Choi, Janghyeok Yoon, Kwangsoo Kim, Jae Yeol Lee, and Cheol Han Kim. SAO network analysis of patents for technology trends identification : A case study of polymer electrolyte membrane technology in proton exchange membrane fuel cells. *Scientometrics*, 88(3), 2011.

[CZZL15] Hongshu Chen, Guangquan Zhang, Donghua Zhu, and Jie Lu. A patent time series processing component for technology intelligence by trend identification functionality. *Neural Computing and Applications*, 26(2), 2015.

[DC12] Cluadia Elena Dinuca and Dumitru Ciobanu. Web Content Mining. *Annals of the University of Petrosani, Economics*, 9(3) :85–92, 2012.

[dCdSF06] Marcio de Miranda Santo, Gilda Massari Coelho, Dalci Maria dos Santos, and Lélio Fellows Filho. Text mining as a valuable tool in foresight exercises : A study on nanotechnology. *Technological Forecasting and Social Change*, 73(8), 2006.

[DCWX10] Xiang Ying Dai, Qing Cai Chen, Xiao Long Wang, and Jun Xu. Online topic detection and tracking of financial news based on hierarchical clustering. In *2010 International Conference on Machine Learning and Cybernetics, ICMLC 2010*, volume 6, 2010.

[DDH90] Scott Deerwester, Susan T. Dumais, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6) :391–407, 1990.

[DFVL14] Rodrigo Dueñas-Fernández, Juan D. Velásquez, and Gaston L'Huillier. Detecting trends on the Web : A multidisciplinary approach. *Information Fusion*, 20(1), 2014.

[DGB+05] Josenildo C. Da Silva, Chris Giannella, Ruchita Bhargava, Hillol Kargupta, and Matthias Klusch. Distributed data mining and agents. *Engineering Applications of Artificial Intelligence*, 2005.

[DHJ+98] George S. Davidson, Bruce Hendrickson, David K. Johnson, Charles E. Meyers, and Brian N. Wylie. Knowledge mining with vxInsight : Discovery through interaction. *Journal of Intelligent Information Systems*, 11(3) :259–285, 1998.

[DKJ18] Ali Dorri, Salil S. Kanhere, and Raja Jurdak. Multi-Agent Systems : A Survey. *IEEE Access*, 2018.

[DS04] George S. Day and Paul Schoemaker. Peripheral Vision : Sensing and Acting on Weak Signals. *Long Range Planning*, 37(2) :117–121, 2004.

[DS05] George S Day and Paul J H Schoemaker. Scanning the Periphery. *Harvard Business Review*, 83(11) :135–148, 2005.

[EJG+19] Olivier Eulaerts, Geraldine Joanny, Jessika Giraldi, Sotirios Fragkiskos, and Sergio Perani. Weak signals in Science and Technologies : 2019 Report. Technical report, 2019.

[EMLS14] R Eckhoff, M Markus, M Lassnig, and Sandra Schön. Detecting weak signals with technologies overview of current technology-enhanced approaches for the detection of weak signals. *International Journal of Trends In Economics, Management Technology*, 3(5) :1–7, 2014.

[EMSS16] Oleg Ena, Nadezhda Mikova, Ozcan Saritas, and Anna Sokolova. A methodology for technology trend monitoring : the case of semantic technologies. *Scientometrics*, 108(3), 2016.

[Faw03] Tom Fawcett. "In vivo" spam filteringg : A challenge problem for KDD, 2003.

[FC08] Teng Kai Fan and Chia Hui Chang. Exploring evolutionary technical trends from academic research papers. In *DAS 2008 - Proceedings of the 8th IAPR International Workshop on Document Analysis Systems*, 2008.

[FD95] Ronen Feldman and Ido Dagan. Knowledge Discovery in Textual Databases (KDT). *International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 112–117, 1995.

[FPSS96] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. From data mining to knowledge discovery in databases. *AI Magazine*, 1996.

[FZYL14] Yixiang Fang, Haijun Zhang, Yunming Ye, and Xutao Li. Detecting hot topics from Twitter : A multiview approach. *Journal of Information Science*, 40(5), 2014.

[GALR16] Gregorio González-Alcaide, Pedro Llorente, and José M. Ramos. Bibliometric indicators to identify emerging research fields : publications on mass gatherings. *Scientometrics*, 109(2), 2016.

[GCPP05] Nicolas Grandjean, Brigitte Charpiot, Carlos Andres Pena, and Manuel C. Peitsch. Competitive intelligence and patent analysis in drug discovery. *Drug Discovery Today : Technologies*, 2(3), 2005.

[GCPT07] Amir Globerson, Gal Chechik, Fernando Pereira, and Professor Naftali Tishby. Euclidean Embedding of Co-occurrence Data. *Journal of Machine Learning Research*, 8 :2265–2295, 2007.

[GJS13] Youngjung Geum, Jeonghwan Jeon, and Hyeonju Seol. Identifying technological opportunities using the novelty detection technique : a case of laser technology in semiconductor manufacturing. *Technology Analysis and Strategic Management*, 25(1), 2013.

[GM12] Jan M. Gerken and Martin G. Moehrle. A new instrument for technology monitoring : Novelty in patents measured by semantic patent analysis, 2012.

[GMP10] V. Gunes, M. Menard, and Simon Petitrenaud. Multiple classifier systems : tools and methods. In C.H. Chen, editor, *Handbook of Pattern Recognition and Computer Vision*, pages 23–46. World Scientific, 2010.

[GT12] Wolfgang Glänzel and Bart Thijs. Using 'core documents' for detecting and labelling new emerging topics. *Scientometrics*, 91(2), 2012.

[GU10] Saurabh Goorha and Lyle Ungar. Discovery of significant emerging trends. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2010.

[GV17] Wolfgang Gaul and Dominique Vincent. Evaluation of the evolution of relationships between topics over time. *Advances in Data Analysis and Classification*, 11(1) :159–178, mar 2017.

[GWB11] Hanning Guo, Scott Weingart, and Katy Börner. Mixed-indicators model for identifying emerging research areas. *Scientometrics*, 89(1), 2011.

[HBLH94] William Hersh, Chris Buckley, T. J. Leone, and David Hickam. 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*, pages 192–201. Association for Computing Machinery, Inc, aug 1994.

[HC14] Mu Hsuan Huang and Chia Pin Chang. Detecting research fronts in OLED field using bibliographic coupling with sliding window. *Scientometrics*, 98(3), 2014.

[HCB13] Hue Cao Hong, Guillaume Chiron, and Alain Boucher. A multi-agent model for image browsing and retrieval. *Studies in Computational Intelligence*, 457, 2013.

[HHWN02] Susan Havre, Elizabeth Hetzler, Paul Whitney, and Lucy Nowell. ThemeRiver : Visualizing thematic changes in large document collections. *IEEE Transactions on Visualization and Computer Graphics*, 8(1) :9–20, 2002.

[Hil06] Elina Hiltunen. Was it a wild card or just our blindness to gradual change ?, 2006.

[Hil07] Elina Hiltunen. The futures window : a medium for presenting visual weak signals to trigger employees' futures thinking in organizations. *Kauppakorkeakoulun julkaisuja. Helsinki*, 2007.

[Hil08] Elina Hiltunen. The future sign and its three dimensions. *Futures*, 40(3) :247–260, 2008.

[HKD13] Abdelhakim Herrouz, Chabane Khentout, and Mahieddine Djoudi. Overview of Web Content Mining Tools. *The International Journal of Engineering And Science (IJES)*, 2(6), 2013.

[HKP12] Jiawei Han, Micheline Kamber, and Jian Pei. *Data Mining : Concepts and Techniques*. 2012.

[HL04] Bryan Horling and Victor Lesser. A survey of multi-agent organizational paradigms. *Knowledge Engineering Review*, 2004.

[hM14] Abdelkader El hadi Mesli. Intégration des agents dans le Data Mining. Technical report, 2014.

[Hof01] Thomas Hofmann. Unsupervised learning by probabilistic Latent Semantic Analysis. *Machine Learning*, 42(1-2) :177–196, 2001.

[HT12] Mari Holopainen and Marja Toivonen. Weak signals : Ansoff today. *Futures*, 44(3) :198–205, 2012.

[HV17] Nagaratna P. Hegde and B. Varija. Data mining and multi-agent integration. *International Journal of Engineering Trends and Technology*, (April), 2017.

[HZM+15] Ying Huang, Yi Zhang, Jing Ma, Alan L Porter, and Xuefeng Wang. Tracing technology evolution pathways by combining patent citation analysis and tech mining. *2015 Portland International Conference on Management of Engineering Technology*, 2015.

[JP17] K. Jayamalini and M. Ponnavaikko. Research on web data mining concepts, techniques and applications. In *2017 International Conference on Algorithms, Methodology, Models and Applications in Emerging Technologies, ICAMMAET 2017*, 2017.

[JPJ12a] Sunghae Jun, Sang Sung Park, and Dong Sik Jang. Patent Management for Technology Forecasting : A Case study of the bio-industry. *Journal of Intellectual Property Rights*, 17(6), 2012.

[JPJ12b] Sunghae Jun, Sang Sung Park, and Dong Sik Jang. Technology forecasting using matrix map and patent clustering. *Industrial Management and Data Systems*, 112(5), 2012.

[JW98] Nicholas R Jennings and Michael J Wooldridge. *Applications of Intelligent Agents*. 1998.

[JY15] Yujin Jeong and Byungun Yoon. Development of patent roadmap based on technology roadmap by analyzing patterns of patent development. *Technovation*, 39-40(1), 2015.

[KGL+14] Byunghoon Kim, Gianluca Gazzola, Jae Min Lee, Dohyun Kim, Kanghoe Kim, and Myong K. Jeong. Inter-cluster connectivity analysis for technology opportunity discovery. *Scientometrics*, 98(3), 2014.

[KGP+04] April Kontostathis, Leon M. Galitsky, William M. Pottenger, Soma Roy, and Daniel J. Phelps. A Survey of Emerging Trend Detection in Textual Data Mining. *Survey of Text Mining*, pages 185–224, 2004.

[KHJJ12] Jinhyung Kim, Myunggwon Hwang, Do Heon Jeong, and Hanmin Jung. Technology trends analysis and forecasting application based on decision tree and statistical feature analysis. *Expert Systems with Applications*, 39(16), 2012.

[KKB+13] Seonho Kim, You Eil Kim, Kuk Jin Bae, Sung Bae Choi, Jong Kyu Park, Young Duk Koo, Young Wook Park, Hyun Kyoo Choi, Hyun Moo Kang, and Sung Wha Hong. NEST : A quantitative model for detecting emerging trends using a global monitoring expert network and bayesian network. *Futures*, 52, 2013.

[KKHR15] Daehoon Kim, Daeyong Kim, Eenjun Hwang, and Seungmin Rho. TwitterTrends : a spatio-temporal trend detection and related keywords recommendation scheme. *Multimedia Systems*, 21(1), 2015.

[KL17] Jieun Kim and Changyong Lee. 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, 2017.

[KLK+15] Namil Kim, Hyeokseong Lee, Wonjoon Kim, Hyunjong Lee, and Jong Hwan Suh. Dynamic patterns of industry convergence : Evidence from a large amount of unstructured data. *Research Policy*, 44(9), 2015.

[KMS06] Tuomo Kakkonen, Niko Myller, and Erkki Sutinen. Applying latent Dirichlet allocation to automatic essay grading. *5th International Conference on NLP, FinTAL 2006 Turku, Finland, August 23-25, 2006 Proceedings*, 2006.

[KMST08] Tuomo Kakkonen, Niko Myller, Erkki Sutinen, and Jari Timonen. Comparison of dimension reduction methods for automated essay grading. *Educational Technology and Society*, 11(3) :275–288, jul 2008.

[Köh16] Arne Köhn. Evaluating Embeddings using Syntax-based Classification Tasks as a Proxy for Parser Performance. pages 67–71. Association for Computational Linguistics (ACL), aug 2016.

[KTKK15] Mirko Kämpf, Eric Tessenow, Dror Y. Kenett, and Jan W. Kantelhardt. The detection of emerging trends using Wikipedia traffic data and context networks. *PLoS ONE*, 10(12), 2015.

[Kuo10] Tuomo Kuosa. Futures signals sense-making framework (FSSF) : A start-up tool to analyse and categorise weak signals, wild cards, drivers, trends and other types of information. *Futures*, 42(1) :42–48, 2010.

[KvdG14] Jonas Keller and Heiko A. von der Gracht. The influence of information and communication technology (ICT) on future foresight processes - Results from a Delphi survey. *Technological Forecasting and Social Change*, 85, 2014.

[LAS97] Brian Lent, Rakesh Agrawal, and Ramakrishnan Srikant. Discovering Trends in Text Databases. *Proc 3rd Int Conf Knowledge Discovery and Data Mining KDD*, pages 227–230, 1997.

[LC04] B Liu and K Chen-Chuan Chang. Editorial : Special Issue on Web Content Mining. *ACM SiIGKDD Explorations Newsletters*, 2004.

[LCA15] Julio Cesar Louzada Pinto, Tijani Chahed, and Eitan Altman. Trend detection in social networks using Hawkes processes. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2015*, 2015.

[LCB12] Jey Han Lau, Nigel Collier, and Timothy Baldwin. On-line trend analysis with topic models : Twitter trends detection topic model online. In *24th International Conference on Computational Linguistics - Proceedings of COLING 2012 : Technical Papers*, 2012.

[Lee08] Woo Hyoung Lee. How to identify emerging research fields using scientometrics : An example in the field of Information Security. *Scientometrics*, 76(3), 2008.

[LG14a] Omer Levy and Yoav Goldberg. Linguistic regularities in sparse and explicit word representations. In *CoNLL 2014 - 18th Conference on Computational Natural Language Learning, Proceedings*, pages 171–180, 2014.

[LG14b] Omer Levy and Yoav Goldberg. Neural word embedding as implicit matrix factorization. In *Advances in Neural Information Processing Systems*, volume 3, pages 2177–2185. Neural information processing systems foundation, 2014.

[LJP11] Changyong Lee, Jeonghwan Jeon, and Yongtae Park. Monitoring trends of technological changes based on the dynamic patent lattice : A modified formal concept analysis approach. *Technological Forecasting and Social Change*, 78(4), 2011.

[LKK10] Jae Yun Lee, Heejung Kim, and Pan Jun Kim. Domain analysis with text mining : Analysis of digital library research trends using profiling methods. *Journal of Information Science*, 36(2), 2010.

[LKS+14] Yongho Lee, So Young Kim, Inseok Song, Yongtae Park, and Juneseuk Shin. Technology opportunity identification customized to the technological capability of SMEs through two-stage patent analysis. *Scientometrics*, 100(1), 2014.

[LPZ15] Jianhong Luo, Xuwei Pan, and Xiyong Zhu. Identifying digital traces for business marketing through topic probabilistic model. *Technology Analysis and Strategic Management*, 27(10), 2015.

[LSLL09] Duen Ren Liu, Meng Jung Shih, Churn Jung Liau, and Chin Hui Lai. Mining the change of event trends for decision support in environmental scanning. *Expert Systems with Applications*, 36(2 PART 1), 2009.

[LZL+13] Yingjie Lu, Pengzhu Zhang, Jingfang Liu, Jia Li, and Shasha Deng. Health-Related Hot Topic Detection in Online Communities Using Text Clustering. *PLoS ONE*, 8(2), 2013.

[Mai19a] Julien Maitre. A Wikipedia dataset of 5 categories. jun 2019.

[Mai19b] Julien Maitre. A Wikipedia dataset of Ebola disease related articles. dec 2019.

[MCC12] Sandro Mendonça, Gustavo Cardoso, and João Caraça. The strategic strength of weak signal analysis. *Futures*, 2012.

[MCCD13] Tomas Mikolov, Greg Corrado, Kai Chen, and Jeffrey Dean. 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, 2013.

[MCKoR04] Sandro Mendonça, Miguel Pine Cunha, Jari Kaivo-oja, and Frank Ruff. Wild cards, weak signals and organisational improvisation. *Futures*, 2004.

[MdFdA+14] Douglas Henrique Milanez, Leandro Innocentini Lopes de Faria, Roniberto Morato do Amaral, Daniel Rodrigo Leiva, and José Angelo Rodrigues Gregolin. Patents in nanotechnology : an analysis using macro-indicators and forecasting curves. *Scientometrics*, 101(2), 2014.

[ME02] Michel Ménard and Michel Eboueya. Extreme physical information and objective function in fuzzy clustering. *Fuzzy Sets and Systems*, 128(3) :285–303, 2002.

[Mén01] Michel Ménard. Fuzzy clustering and switching regression models using ambiguity and distance rejects. *Fuzzy Sets and Systems*, 122(3) :363–399, 2001.

[MG18] Christian Mühlroth and Michael Grottke. A systematic literature review of mining weak signals and trends for corporate foresight. *Journal of Business Economics*, 88(5), 2018.

[MHCS15] A. L. M. Moreira, T. W. N. Hayashi, G. P. Coelho, and A. E. A. Silva. A Clustering Method for Weak Signals to Support Anticipative Intelligence. *International Journal of Artificial Intelligence and Expert Systems*, 6(1), 2015.

[MHKB16] O. Mryglod, Yu Holovatch, R. Kenna, and B. Berche. Quantifying the evolution of a scientific topic : reaction of the academic community to the Chornobyl disaster. *Scientometrics*, 106(3), 2016.

[MK10] Michael Mathioudakis and Nick Koudas. TwitterMonitor : Trend Detection over the Twitter Stream. *SIGMOD '10 Proceedings of the 2010 ACM SIGMOD International Conference on Management of data*, pages 1155–1158, 2010.

[MP15] Jing Ma and Alan L. Porter. Analyzing patent topical information to identify technology pathways and potential opportunities. *Scientometrics*, 102(1), 2015.

[MR09] Oded Maimon and Lior Rokach. Introduction to Knowledge Discovery and Data Mining. In *Data Mining and Knowledge Discovery Handbook*. 2009.

[MSC+13] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems*. Neural information processing systems foundation, 2013.

[MWCE07] Andrew McCallum, Xuerui Wang, and Andrés Corrada-Emmanuel. Topic and role discovery in social networks with experiments on enron and academic email. *Journal of Artificial Intelligence Research*, 2007.

[MZ05] Qiaozhu Mei and Cheng Xiang Zhai. Discovering evolutionary theme patterns from text - An exploration of Temporal Text Mining. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2005.

[NAXC08] Ramesh M Nallapati, Amr Ahmed, Eric P Xing, and William W Cohen. Joint latent topic models for text and citations. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 542–550, 2008.

[NRC06] Olfa Nasraoui, Carlos Rojas, and Cesar Cardona. A framework for mining evolving trends in Web data streams using dynamic learning and retrospective validation. *Computer Networks*, 50(10), 2006.

[NSL16] Heeyong Noh, Young Keun Song, and Sungjoo Lee. Identifying emerging core technologies for the future : Case study of patents published by leading telecommunication organizations. *Telecommunications Policy*, 40(10-11), 2016.

[NSY16] Khanh Ly Nguyen, Byung Joo Shin, and Seong Joon Yoo. Hot topic detection and technology trend tracking for patents utilizing term frequency and proportional document frequency and semantic information. In *2016 International Conference on Big Data and Smart Computing, BigComp 2016*, 2016.

[PD95] Alan L. Porter and Michael J. Detampel. Technology opportunities analysis. *Technological Forecasting and Social Change*, 49(3) :237–255, 1995.

[PKB+11] Heum Park, Eunsun Kim, Kuk Jin Bae, Hyuk Hahn, Tae Eung Sung, and Hyuk Chul Kwon. Detection and analysis of trend topics for global scientific literature using feature selection based on Gini-Index. In *Proceedings - International Conference on Tools with Artificial Intelligence, ICTAI*, 2011.

[PKL+16] Sangsung Park, Juhwan Kim, Hongchul Lee, Dongsik Jang, and Sunghae Jun. Methodology of technological evolution for three-dimensional printing. *Industrial Management and Data Systems*, 116(1), 2016.

[PKM01] William M. Pottenger, Yong-Bin Kim, and Daryl D. Meling. HDDI^{TM} : Hierarchical Distributed Dynamic Indexing. pages 319–333, 2001.

[PKS15] Siddhant Patil, Sayali Karnik, and Vinaya Sawant. A Review on Multi-Agent Data Mining Systems. *International Journal of Computer Science and Information Technologies*, 6(6) :4888–4893, 2015.

[PNLM13] Nina Preschitschek, Helen Niemann, Jens Leker, and Martin G. Moehrle. Anticipating industry convergence : Semantic analyses vs IPC co-classification analyses of patents. *Foresight*, 15(6), 2013.

[PNML12] Nina Preschitschek, Helen Niemann, Martin G. Moehrle, and Jens Leker. Semantic analyses vs. IPC Co-classification analyses of patents : Which one better serves to anticipate converging industries ? In *2012 Proceedings of Portland International Center for Management of Engineering and Technology : Technology Management for Emerging Technologies, PICMET'12*, 2012.

[PVO13] Marco A. Palomino, Alexandra Vincenti, and Richard Owen. Optimising web-based information retrieval methods for horizon scanning. *foresight*, 15(3), 2013.

[PWY+13] Jon Parker, Yifang Wei, Andrew Yates, Ophir Frieder, and Nazli Goharian. A framework for detecting public health trends with Twitter. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2013*, 2013.

[RB12] René Rohrbeck and Manuel Bade. Environmental Scanning, Futures Research, Strategic Foresight and Organizational Future Orientation : A Review, Integration, and Future Research Directions. pages 1–14, 2012.

[RCK21] Pauline Rousseau, Daniel Camara, and Dimitris Kotzinos. Weak signal detection and identification in large data sets : a review of methods and applications, 2021.

[RCY06] Loîs Rigouste, Olivier Cappé, and François Yvon. Quelques observations sur le modèle LDA. *Journées internationales d'Analyse statistique des Données Textuelles*, 8 :819–830, 2006.

[RGP02] Soma Roy, David Gevry, and William M. Pottenger. Methodologies for trend detection in textual data mining. -, -(-) :–, 2002.

[RHL09] Stefan Romberg, Eva Hörster, and Rainer Lienhart. Multimodal pLSA on visual features and tags. In *Proceedings - 2009 IEEE International Conference on Multimedia and Expo, ICME 2009*, 2009.

[RN10] Stuart Russell and Peter Norvig. *Artificial Intelligence A Modern Approach Third Edition.* 2010.

[RRSZ14] Sven Rill, Dirk Reinel, Jörg Scheidt, and Roberto V. Zicari. PoliTwi : Early detection of emerging political topics on twitter and the impact on concept-level sentiment analysis. *Knowledge-Based Systems*, 69(1), 2014.

[RTK+16] Andrew Rodriguez, Ali Tosyali, Byunghoon Kim, Jeongsub Choi, Jae Min Lee, Byoung Youl Coh, and Myong K. Jeong. Patent Clustering and Outlier Ranking Methodologies for Attributed Patent Citation Networks for Technology Opportunity Discovery. *IEEE Transactions on Engineering Management*, 63(4), 2016.

[RVR10] Vuda Sreenivasa Rao, S Vidyavathi, and G Ramaswamy. Distributed Data Mining and Agent Mining Interaction and Integration : A Novel Approach. Technical report, 2010.

[SA00] Russell Swan and James Allan. Automatic generation of overview timelines. *SIGIR Forum (ACM Special Interest Group on Information Retrieval)*, pages 49–56, 2000.

[SBMN13] Richard Socher, John Bauer, Christopher D. Manning, and Andrew Y. Ng. Parsing with compositional vector grammars. In *ACL 2013 - 51st Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, volume 1, pages 455–465. Association for Computational Linguistics (ACL), 2013.

[Sch16] Christof Schöch. A word2vec model file built from the French Wikipedia XML Dump using gensim. oct 2016.

[SJ00] Russell Swan and David Jensen. TimeMines : Constructing Timelines with Statistical Models of Word Usage. *ACM SIGKDD 2000 Workshop on Text Mining*, pages 73–80, 2000.

[SKJ14] Min Song, Meen Chul Kim, and Yoo Kyung Jeong. Analyzing the political landscape of 2012 korean presidential election in twitter. *IEEE Intelligent Systems*, 29(2), 2014.

[SKT+11] Naoki Shibata, Yuya Kajikawa, Yoshiyuki Takeda, Ichiro Sakata, and Katsumori Matsushima. Detecting emerging research fronts in regenerative medicine by the citation network analysis of scientific publications. *Technological Forecasting and Social Change*, 78(2), 2011.

[SLH10] Meng Jung Shih, Duen Ren Liu, and Ming Li Hsu. Discovering competitive intelligence by mining changes in patent trends. *Expert Systems with Applications*, 37(4), 2010.

[SLMJ15] Tobias Schnabel, Igor Labutov, David Mimno, and Thorsten Joachims. Evaluation methods for unsupervised word embeddings. In *Conference Proceedings - EMNLP 2015 : Conference on Empirical Methods in Natural Language Processing*, pages 298–307, 2015.

[SPW+13] Richard Socher, Alex Perelygin, Jean Y. Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. 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*, pages 1631–1642. Association for Computational Linguistics (ACL), 2013.

[SS11] Ozcan Saritas and Jack E. Smith. The Big Picture - trends, drivers, wild cards, discontinuities and weak signals. *Futures*, 43(3) :292–312, 2011.

[SSS08] Zhi Yong Shen, Jun Sun, and Yi Dong Shen. Collective latent dirichlet allocation. In *Proceedings - IEEE International Conference on Data Mining, ICDM*, pages 1019–1024, 2008.

[SSW+17] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. Fake news detection on social media : A data mining perspective, 2017.

[TBG+14] Yulia Tsvetkov, Leonid Boytsov, Anatole Gershman, Eric Nyberg, and Chris Dyer. 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, pages 248–258, 2014.

[TFL+15] Yulia Tsvetkov, Manaal Faruqui, Wang Ling, Guillaume Lample, and Chris Dyer. Evaluation of word vector representations by subspace alignment. In *Conference Proceedings - EMNLP 2015 : Conference on Empirical Methods in Natural Language Processing*, pages 2049–2054, 2015.

[TH16] Yi Ning Tu and Shu Lan Hsu. Constructing conceptual trajectory maps to trace the development of research fields. *Journal of the Association for Information Science and Technology*, 67(8), 2016.

[THF03] Quan Thanh Tho, Siu Cheung Hui, and Alvis Fong. Web mining for identifying research trends. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2911 :290–301, 2003.

[TRB10] Joseph Turian, Lev Ratinov, and Yoshua Bengio. 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*, pages 384–394, 2010.

[TS12] Yi Ning Tu and Jia Lang Seng. Indices of novelty for emerging topic detection. *Information Processing and Management*, 48(2), 2012.

[TSV14] Dirk Thorleuchter, Tobias Scheja, and Dirk Van Den Poel. Semantic weak signal tracing. *Expert Systems with Applications*, 41(11) :5009–5016, 2014.

[TTY14] Toshimitsu Takahashi, Ryota Tomioka, and Kenji Yamanishi. Discovering emerging topics in social streams via link-anomaly detection. *IEEE Transactions on Knowledge and Data Engineering*, 26(1), 2014.

[TV13] Dirk Thorleuchter and Dirk Van Den Poel. Weak signal identification with semantic web mining. *Expert Systems with Applications*, 40(12) :4978–4985, 2013.

[TV15] D. Thorleuchter and D. Van den Poel. Idea mining for web-based weak signal detection. *Futures*, 66 :25–34, 2015.

[TWTDT11] Charles V. Trappey, Hsin Ying Wu, Fataneh Taghaboni-Dutta, and Amy J.C. Trappey. Using patent data for technology forecasting : China RFID patent analysis. *Advanced Engineering Informatics*, 25(1), 2011.

[VBV10] Mark Veugelers, Jo Bury, and Stijn Viaene. Linking technology intelligence to open innovation. *Technological Forecasting and Social Change*, 77(2), 2010.

[vdGVD10] Heiko A. von der Gracht, Christoph Robert Vennemann, and Inga Lena Darkow. Corporate foresight and innovation management : A portfolio-approach in evaluating organizational development. *Futures*, 42(4), 2010.

[VH08] Laurens Van Der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9 :2579–2625, nov 2008.

[vVR21] Barbara L. van Veen and J. Roland Ortt. Unifying Weak Signals Definitions to Improve Construct Understanding. *Futures*, 134(August) :102837, 2021.

[WCK10] Ming Yeu Wang, Dong Shang Chang, and Chih Hsi Kao. Identifying technology trends for R and D planning using TRIZ and text mining. *R and D Management*, 40(5), 2010.

[WCL14] Xiaoguang Wang, Qikai Cheng, and Wei Lu. Analyzing evolution of research topics with NEViewer : a new method based on dynamic co-word networks. *Scientometrics*, 101(2), 2014.

[WHM09] Wei Lee Woon, Andreas Henschel, and Stuart Madnick. A framework for technology forecasting and visualization. In *2009 International Conference on Innovations in Information Technology, IIT '09*, 2009.

[WLT+15] Jing Wang, Li Li, Feng Tan, Ying Zhu, and Weisi Feng. Detecting hotspot information using multi-attribute based topic model. *PLoS ONE*, 10(10), 2015.

[WM06] Xuerui Wang and Andrew McCallum. Topics over Time : A non-markov continuous-time model of topical trends. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, volume 2006, 2006.

[WM12] Wei Lee Woon and Stuart Madnick. Semantic distances for technology landscape visualization. *Journal of Intelligent Information Systems*, 39(1), 2012.

[WQZ+15] Xuefeng Wang, Pengjun Qiu, Donghua Zhu, Liliana Mitkova, Ming Lei, and Alan L. Porter. Identification of technology development trends based on subject-action-object analysis : The case of dye-sensitized solar cells. *Technological Forecasting and Social Change*, 98, 2015.

[WRP+13] Tamar C. Weenen, Bahar Ramezanpour, Esther S. Pronker, Harry Commandeur, and Eric Claassen. Food-pharma convergence in medical nutrition - Best of both worlds ? *PLoS ONE*, 8(12), 2013.

[WSLS10] Feng Shang Wu, Chun Chi Shiu, Pei Chun Lee, and Hsin Ning Su. Integrated methodologies for mapping and forecasting science and technology trends : A case

of etching technology. In *PICMET '10 - Portland International Center for Management of Engineering and Technology, Proceedings - Technology Management for Global Economic Growth*, 2010.

[WWC08] H Wu, Y J Wang, and X Cheng. Incremental Probabilistic Latent Semantic Analysis for Automatic Question Recommendation. *Recsys'08 : Proceedings of the 2008 Acm Conference on Recommender Systems*, 2008.

[WZB10] Robert Wetzker, Carsten Zimmermann, and Christian Bauckhage. Detecting trends in social bookmarking systems : A del.icio.us endeavor. In *International Journal of Data Warehousing and Mining*, volume 6, 2010.

[XZJ+16] Wei Xie, Feida Zhu, Jing Jiang, Ee Peng Lim, and Ke Wang. TopicSketch : Real-time bursty topic detection from twitter. *IEEE Transactions on Knowledge and Data Engineering*, 28(8), 2016.

[YK12] Janghyeok Yoon and Kwangsoo Kim. Detecting signals of new technological opportunities using semantic patent analysis and outlier detection. *Scientometrics*, 2012.

[YLLL16] Liang Yang, Hongfei Lin, Yuan Lin, and Shengbo Liu. Detection and Extraction of Hot Topics on Chinese Microblogs. *Cognitive Computation*, 8(4), 2016.

[Yoo12] Janghyeok Yoon. Detecting weak signals for long-term business opportunities using text mining of Web news. *Expert Systems with Applications*, 39(16) :12543–12550, 2012.

[YP07] Byungun Yoon and Yongtae Park. Development of new technology forecasting algorithm : Hybrid approach for morphology analysis and conjoint analysis of patent information. *IEEE Transactions on Engineering Management*, 2007.

[YPC98] Yiming Yang, Tom Pierce, and Jaime Carbonell. Study on retrospective and on-line event detection. *SIGIR Forum (ACM Special Interest Group on Information Retrieval)*, pages 28–36, 1998.

[ZCP+15] Weizhong Zhao, James J. Chen, Roger Perkins, Zhichao Liu, Weigong Ge, Yijun Ding, and Wen Zou. A heuristic approach to determine an appropriate number of topics in topic modeling. *BMC Bioinformatics*, 16(13), sep 2015.

Détection et analyse des signaux faibles. Développement d'un framework d'investigation numérique pour un service caché Lanceurs d'alerte.

Ce manuscrit 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. Nous proposons une chaine d'extraction à partir de corpus de document, d'analyse semi-automatisée et de recherche au moyen de requêtes Web pour *in fine*, proposer des tableaux de bord décrivant les *signaux faibles* potentiels. Nous identifions et levons un certain nombre de verrous méthodologiques et technologiques inhérents : 1) à l'analyse automatique de contenus textuels avec un minimum d'*a priori*, 2) à l'enrichissement de l'information à partir de recherches Web 3) à la visualisation sous forme de tableau de bord et d'une représentation dans un espace 3D interactif. Ces approches, statique et dynamique, sont appliquées au contexte du data journalisme, et en particulier, au traitement, analyse et hiérarchisation d'informations hétérogènes présentes dans des documents. Cette thèse propose également une étude de faisabilité et de prototypage par la mise en œuvre d'une chaine de traitement sous forme d'un logiciel. La construction de celui-ci a nécessité la caractérisation d'un *signal faible* pour lequel nous avons proposé une définition. Notre objectif est de fournir un outil paramétrable et générique à toute thématique.

La solution que nous proposons repose sur deux approches : statique et dynamique. Dans l'approche statique, contrairement aux approches existantes nécessitant la connaissance de termes pertinents dans un domaine spécifique, nous proposons une solution s'appuyant sur des techniques nécessitant une intervention moindre de l'expert du domaine. Dans ce contexte, nous proposons une nouvelle approche de modélisation thématique multi-niveaux. Cette méthode d'approche conjointe combine une modélisation thématique, un plongement de mots et un algorithme où le recours à un expert du domaine permet d'évaluer la pertinence des résultats et d'identifier les thèmes porteurs de *signaux faibles* potentiels. Dans l'approche dynamique, nous intégrons une solution de veille à partir des *signaux faibles* potentiels trouvées dans les corpus initiaux et effectuons un suivi pour étudier leur évolution. Nous proposons donc une solution d'*agent mining* combinant *data mining* et système multi-agents où des agents animés par des forces d'attraction/répulsion représentant documents et mots se déplacent. La visualisation des résultats est réalisée sous forme de tableau de bord et de représentation dans un espace 3D interactif dans un client Unity. Dans un premier temps, l'approche statique a été évaluée dans une preuve de concept sur des corpus synthétiques et réelles utilisés comme vérité terrain. L'ensemble de la chaine de traitement (approches statique et dynamique), mise en œuvre dans le logiciel WILD, est dans un deuxième temps appliquée sur des données réelles provenant de bases documentaires.

Mots clés : modélisation thématique, plongement de mots, LDA, Word2Vec, distance de Bhattacharyya, tf-idf, agent mining, data mining, système multi-agents, signal faible, data journalisme.

Detection and analysis of weak signals. Development of a digital investigation framework for a hidden whistleblower service.

This manuscript provides the basis for a complete chain of document analysis for a whistleblower service, such as GlobalLeaks. We propose a chain of semi-automated analysis of text document and search using web search queries to *in fine* present dashboards describing weak signals. We identify and solve methodological and technological barriers inherent to : 1) automated analysis of text document with minimum *a priori* information, 2) enrichment of information using web search 3) data visualization dashboard and 3D interactive environment. These static and dynamic approaches are used in the context of data journalism for processing heterogeneous types of information within documents. This thesis also proposed a feasibility study and prototyping by the implementation of a processing chain in the form of a software. This construction requires a *weak signal* definition. Our goal is to provide configurable and generic tool.

Our solution is based on two approaches : static and dynamic. In the static approach, we propose a solution requiring less intervention from the domain expert. In this context, we propose a new approach of multi-level topic modeling. This joint approach combines topic modeling, word embedding and an algorithm. The use of a expert helps to assess the relevance of the results and to identify topics with *weak signals*. In the dynamic approach, we integrate a solution for monitoring *weak signals* and we follow up to study their evolution. We therefore propose and agent mining solution which combines data mining and multi-agent system where agents representing documents and words are animated by attraction/repulsion forces. The results are presented in a data visualization dashboard and a 3D interactive environment in Unity. First, the static approach is evaluated in a proof-of-concept with synthetic and real text corpus. Second, the complete chain of document analysis (static and dynamic) is implemented in a software and are applied to data from document databases.

Keywords : topic modeling, word embedding, LDA, Word2Vec, Bhattacharyya distance, tf-idf, agent mining, data mining, multi-agent system, weak signal, data journalism.

L3i - Laboratoire Informatique,
Image et Interaction
Avenue Michel Crépeau
17042 La Rochelle Cedex 01