Appariement de contenus textuels dans le domaine de la presse en ligne : Développement et adaptation d'un système de recherche d'information

Adèle Désoyer

Thèse présentée pour obtenir le grade de docteur

Université Paris Nanterre

❧

Laboratoire MoDyCo - UMR 7114
École doctorale 139 "Connaissance, Langage, Modélisation"

Discipline : Sciences du Langage
Spécialité : Traitement Automatique des Langues

Appariement de contenus textuels dans le domaine de la presse en ligne : Développement et adaptation d'un système de recherche d'information

---

Par : Adèle Désoyer

Sous la direction de Delphine Battistelli

Membres du jury :

Rapporteure : Brigitte Grau, Professeure des universités, ENSIIE, LIMSI
Rapporteur : Ludovic Tanguy, Maître de conférences – HDR, Univ. Toulouse 2, CLLE-ERSS
Examinatrice : Haïfa Zargayouna, Maître de conférences, Univ. Paris 13, LIPN
Examinateur : Jean-Luc Minel, Professeur des universités, Univ. Paris Nanterre, MoDyCo
Directrice : Delphine Battistelli, Professeure des universités, Univ. Paris Nanterre, MoDyCo
Invité : Yann Battard, Co-fondateur, MEDIABONG

Date de soutenance : 27 novembre 2017

Remerciements

Je tiens à remercier en premier lieu les personnes sans qui ce travail n'aurait jamais vu le jour. Merci à ma directrice de recherche, Delphine Battistelli, qui m'a offert l'opportunité de réaliser cette thèse et dont l'encadrement et les remarques toujours pertinentes ont permis à ce travail d'aboutir. Merci également à Yann Battard et Laurent Bury, fondateurs de Mediabong, qui m'ont accordé leur confiance dès le début de cette aventure et donné un environnement propice à mener à bien mes travaux de recherche.

Je remercie aussi vivement Brigitte Grau et Ludovic Tanguy pour avoir accepté de rapporter ce travail et pour leurs remarques et conseils qui ont contribué à l'améliorer. Merci également à Haïfa Zargayouna pour sa participation au jury de soutenance, de même que Jean-Luc Minel que je remercie par ailleurs pour l'intérêt qu'il a accordé à mes travaux qui n'auraient pu progresser sans son soutien.

Mes remerciements les plus sincères vont également à l'ensemble des collaborateurs de Mediabong qui ont fait de cette première expérience professionnelle un moment dont je me souviendrai toujours. Merci de nouveau à Yann et à Laurent qui m'ont accueillie au sein de leur entreprise et m'ont laissé toute l'autonomie et l'initiative de faire ce qui me semblait juste. Merci à Michaël sans qui je n'aurais su par où commencer à mon arrivée, ainsi que pour l'ensemble de ses contributions à ce travail. Merci à Khalid pour sa collaboration sur ce projet et pour tous les enseignements qu'il a pu m'apporter. Merci à Pierre et à Vivien pour leur réactivité face aux problèmes techniques, notamment ceux que j'ai pu engendrer lors d'expériences malhabiles... Merci bien sûr à ceux qui ont consacré une partie de leur temps à l'annotation des données qui fondent ce travail de recherche : Clémentine, Yann, rien n'aurait avancé sans votre investissement et votre patience. Merci infiniment à Geoffray pour la relecture attentive de ce mémoire et pour ses commentaires qui m'ont aidée à l'enrichir. Merci à tous les autres que je ne cite pas exhaustivement mais qui ont aussi apporté, par leur conversation, leur humour et leurs savoirs, une pierre à l'édifice.

Ces remerciements ne seraient pas complets sans que n'y figurent mes collègues et désormais amis de MoDyCo. Merci à Elise pour l'attention qu'elle a accordée à ma thèse, pour ses remarques et conseils qui ont été plus que précieux à un moment où il m'était difficile d'avancer. Merci à Ilaine dont le travail soigné et réfléchi m'a toujours incitée à approfondir le mien. Un grand merci à toutes les deux, *ma famiglia*, pour vos encouragements qui m'ont permis de mener cette thèse à son terme et pour votre présence indéfectible

au laboratoire qui a souvent été le moteur de la mienne. *Grazie mille a* Veronica pour sa force de caractère et sa volonté qui m'ont aussi encouragée à persévérer jusqu'au bout. Merci bien sûr à Guillaume dont la folie, l'humour et les talents d'imitation m'ont plus d'une fois fait rire aux larmes, m'évitant quelques *journées dans le noir*. Merci enfin à tous les autres, doctorants et chercheurs, qui ont participé à créer une vie de laboratoire dans laquelle je me suis épanouie.

Je dédie ces dernières lignes aux plus proches de mes proches qui ont contribué bien plus qu'ils ne le soupçonnent à la réalisation de ce travail. Merci à mes parents qui m'ont toujours laissée libre de mes choix et soutenue dans ceux que je faisais. Merci à ma sœur, Mathilde, pour sa présence à mes côtés tout au long de cette thèse mais plus encore pour sa présence dans ma vie sans laquelle rien ne serait pareil.

Résumé

L'objectif de cette thèse, menée dans un cadre industriel, est d'apparier des contenus textuels médiatiques. Plus précisément, il s'agit d'apparier à des articles de presse en ligne des vidéos pertinentes, pour lesquelles nous disposons d'une description textuelle. Notre problématique relève donc exclusivement de l'analyse de matériaux textuels, et ne fait intervenir aucune analyse d'image ni de langue orale. Surviennent alors des questions relatives à la façon de comparer des objets textuels, ainsi qu'aux critères mobilisés pour estimer leur degré de similarité. L'un de ces éléments est selon nous la similarité thématique de leurs contenus, autrement dit le fait que deux documents doivent relater le même sujet pour former une paire pertinente. Ces problématiques relèvent du domaine de la recherche d'information (RI), dans lequel nous nous ancrons principalement. Par ailleurs, lorsque l'on traite des contenus d'actualité, la dimension temporelle est aussi primordiale et les problématiques qui l'entourent relèvent de travaux ayant trait au domaine du *Topic Detection and Tracking* (TDT) dans lequel nous nous inscrivons également.

Le système d'appariement développé dans cette thèse distingue donc différentes étapes qui se complètent. Dans un premier temps, l'indexation des contenus fait appel à des méthodes de Traitement Automatique des Langues (TAL) pour dépasser la représentation classique des textes en *sacs de mots*. Ensuite, deux scores sont calculés pour rendre compte du degré de similarité entre deux contenus : l'un relatif à leur similarité thématique, basé sur un modèle vectoriel de RI ; l'autre à leur proximité temporelle, basé sur une fonction empirique. Finalement, un modèle de classification appris à partir de paires de documents, décrites par ces deux scores et annotées manuellement, permet d'ordonnancer les résultats.

L'évaluation des performances du système a elle aussi fait l'objet de questionnements dans ces travaux de thèse. Les contraintes imposées par les données traitées et le besoin particulier de l'entreprise partenaire nous ont en effet contraint à adopter une alternative au protocole classique d'évaluation en RI, le paradigme de Cranfield.

Mots-clés : Système de recherche d'information ; *Topic Detection and Tracking* ; Recommandation basée sur le contenu ; Apprentissage supervisé ; Cadre d'évaluation

Abstract

The goal of this thesis, conducted within an industrial framework, is to pair textual media content. Specifically, the aim is to pair on-line news articles to relevant videos for which we have a textual description. The main issue is then a matter of textual analysis, no image or spoken language analysis was undertaken in the present study. The question that arises is how to compare these particular objects, the texts, and also what criteria to use in order to estimate their degree of similarity. We consider that one of these criteria is the topic similarity of their content, in other words, the fact that two documents have to deal with the same topic to form a relevant pair. This problem fall within the field of Information Retrieval (IR) which is the main strategy called upon in this research. Furthermore, when dealing with news content, the time dimension is of prime importance. To address this aspect, the field of Topic Detection and Tracking (TDT) will also be explored.

The pairing system developed in this thesis distinguishes different steps which complement one another. In the first step, the system uses Natural Language Processing (NLP) methods to index both articles and videos, in order to overcome the traditionnal *bag-of-words* representation of texts. In the second step, two scores are calculated for an article-video pair : the first one reflects their topical similarity and is based on a vector space model ; the second one expresses their proximity in time, based on an empirical function. At the end of the algorithm, a classification model learned from manually annotated document pairs is used to rank the results.

Evaluation of the system's performances raised some further questions in this doctoral research. The constraints imposed both by the data and the specific need of the partner company led us to adapt the evaluation protocol traditionnal used in IR, namely the Cranfield paradigm. We therefore propose an alternative solution for evaluating the system that takes all our constraints into account.

Keywords : Information Retrieval System ; Topic Detection and Tracking ; Content-based Recommendation ; Supervised Learning ; Evaluation framework

Table des matières

Liste des tableaux

Table des figures

Liste des abréviations

EMM	Expression Multi-Mots
EN	Entité Nommée
IDF	*Inverse Document Frequency*
PoS	*Part-of-Speech*
RI	Recherche d'Information
SN	Syntagme Nominal
SRI	Système de Recherche d'Information
SVM	*Support Vector Machine*
TAL	Traitement Automatique des Langues
TDT	*Topic Detection and Tracking*
TF	*Term Frequency*
VSM	*Vector Space Model*

Introduction

Étymologiquement, l'information est ce qui donne forme à l'esprit, en opposition à l'instruction qui lui confère une structure (Littré et al., 1869). Son concept occupe depuis toujours une place centrale dans les sociétés qui se construisent et particulièrement dans l'actuelle que l'on nomme très justement *société de l'information*, où elle se trouve fortement liée à la notion de communication.

La révolution numérique a offert à l'information de nouvelles dimensions par la production constante de données et de là conféré des enjeux cruciaux qui façonnent de nouveaux modes de pensée. Une donnée devient en effet information lorsqu'il est possible de lui donner du sens, de lui faire dire quelque chose. La donnée est devenue l'élément central de cette nouvelle ère, sur laquelle se fondent de nouveaux projets économiques, politiques et plus largement sociétaux pour prédire des modèles qui se veulent robustes puisque basés sur des connaissances réelles de l'existant. Ce que l'on calculait avant à partir d'échantillons de population, dont la représentativité pouvait être mise en cause, se fait aujourd'hui sur les données de tout utilisateur du Web dont un simple clic peut être traduit en trait de personnalité. Désormais, les algorithmes déduisent des informations et construisent des représentations du monde à partir de ce qu'ils observent de la masse de données fournies en entrée : si un individu avec telles caractéristiques a pris telle décision en réponse à tel type de tâche, alors un individu aux caractéristiques similaires devrait prendre la même décision dans le même contexte. Si l'idée est simple, les méthodes de calculs sont bien plus délicates notamment lorsqu'il s'agit de traiter des données complexes, de part leur structure par exemple, comme c'est le cas des contenus textuels qui constituent l'objet d'étude de ce travail.

Une autre dimension d'analyse concerne le fait que l'information seule n'acquiert sa valeur que si elle est transmise, diffusée par ses détenteurs. Véhiculer la connaissance, l'information, c'est participer à la construction de savoirs partagés, sur lesquels reposent les individus et les communautés dans lesquelles ils s'inscrivent. Dans ce paradigme, la communication de l'information est un enjeu de taille dont les médias se sont emparés pour produire en permanence des contenus dont les lecteurs - auditeurs - téléspectateurs - internautes se nourrissent. Cette machine médiatique est celle que nous étudions ici, au travers de ses divers contenus et particulièrement ceux issus de son plus récent canal de diffusion, le Web.

Par ailleurs, si l'information fait partie intégrante du monde actuel, la recherche d'in-

formation est alors une activité décisive pour les individus qui le peuplent. Les tâches relatives à ce domaine cherchent à répondre automatiquement au besoin d'information d'un utilisateur, en recherchant parmi un ensemble de documents ceux qui sont pertinents pour ce besoin. Ces tâches fondamentales, au cœur de nos recherches, ont évolué avec l'essor d'Internet et du *Big Data*. Elles tentent de répondre à de nouvelles problématiques qui sont aussi les nôtres :

— Comment saisir et modéliser le besoin initialement exprimé par l'utilisateur, afin de le satisfaire pleinement ?
— Comment retrouver une information pertinente en réponse à un besoin d'information, au milieu d'une masse de données en constante évolution ?
— Comment représenter les contenus de manière optimale afin de faciliter le processus de recherche d'information ?
— Comment évaluer la réponse du système au regard du besoin d'information initialement exprimé ?

Contexte industriel

Cette thèse est née de la sollicitation du laboratoire MoDyCo par l'entreprise Mediabong, pour travailler ensemble sur un projet au cœur de leur métier : l'association de contenus médiatiques.

Mediabong est une place de marché mettant en relation des producteurs, diffuseurs et annonceurs, proposant l'indexation de contenus vidéos puis leur syndication sur des sites web. Les diffuseurs sont des sites de presse en ligne qui souhaitent intégrer des vidéos dans leurs pages d'article, qui proviennent elles des producteurs. Les annonceurs organisent quant à eux des opérations publicitaires, notamment des campagnes de spots vidéos, que l'on peut régulièrement voir se lancer en amont de la lecture d'une vidéo de contenu sur Internet.

Le revenu de chacun des acteurs – producteur, diffuseur et opérateur [1] – correspond au tiers du revenu global engendré par la publicité précédant la lecture de vidéo éditoriale proposée aux internautes. La promesse de contextualisation vidéo dans les pages séduit annonceurs et agences qui comptent sur cette capacité de ciblage pour maximiser le nombre de vues de vidéos et donc le nombre de consommateurs potentiels du produit promu.

La Figure 1 présente schématiquement le fonctionnement de cette place de marché : lorsqu'une page d'un des partenaires diffuseurs est reçue, son URL est envoyée à Mediabong qui propose en sortie une vidéo contextuelle (*i.e.* dont le contenu illustre ou

1. *i.e.* Mediabong

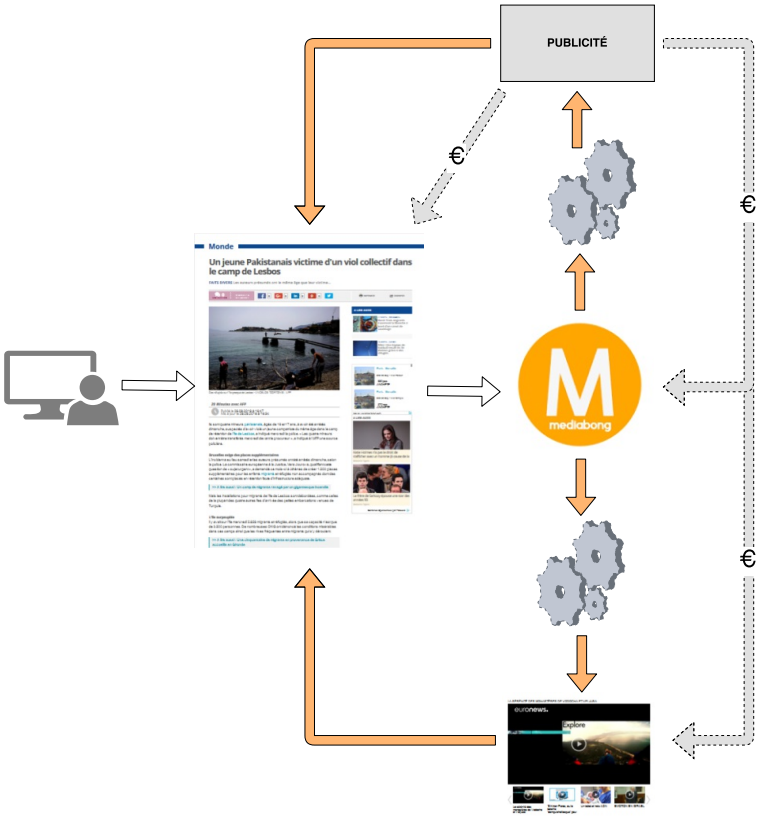

FIGURE 1 – Schéma du fonctionnement de la place de marché MEDIABONG

corrobore celui de l'article). Celle-ci est intégrée en bas d'article et à chaque nouvelle visite de la page par un internaute, deux requêtes sont envoyées : la première au serveur du diffuseur, qui renvoie l'article avec la vidéo sélectionnée ; la seconde à un *adserver* dont la publicité renvoyée sous forme de spot se lance avant la vidéo éditoriale.

En fonction de l'actualité médiatique et de la fluctuation des partenariats diffuseurs, ce sont plusieurs centaines, si ce n'est quelques milliers d'articles qui sont quotidiennement traités par MEDIABONG. Ils nous arrivent sous la forme d'URLs, desquelles on extrait le contenu textuel utile à leur représentation, à savoir le titre, la description et la date de publication. Parallèlement et selon les mêmes variables, plusieurs centaines de nouvelles vidéos viennent quotidiennement alimenter la collection de l'entreprise. Notons ici que ces contenus audiovisuels sont accompagnés d'une description textuelle semi-structurée, sous la forme d'un titre, d'une description et parfois d'une liste de mots-clés [2]. C'est sur cette description textuelle de contenu vidéo que nous travaillons exclusivement, aucune analyse d'image ni de son n'est donc impliquée dans ces travaux de recherche. Toutefois, afin de faciliter la lecture du mémoire, **nous désignerons par la suite ces descriptions textuelles associées aux vidéos sous le seul terme de *vidéos***.

2. Les Annexes A et B présentent respectivement un échantillon d'articles et de vidéos issues de notre corpus, dans lesquelles on peut observer la diversité des thématiques qu'ils recouvrent et des tailles qu'ils font

Notre contribution se situe au niveau du système d'appariement de vidéos aux articles soumis à Mediabong. Avant que ne débute ce projet de thèse, l'entreprise interrogeait un système externe pour rechercher automatiquement des vidéos pour les articles. Mais devant la faible performance de celui-ci, il s'agissait essentiellement d'un traitement manuel opéré par une équipe de quelques salariés en charge de la vérification des sorties du système. Par ailleurs, lorsque le système ne proposait aucun résultat, et ce dans presque 50% des cas, les salariés devaient procéder à une recherche de vidéo manuelle, ce qui était encore davantage coûteux en temps.

La problématique initialement formulée était donc de développer un système capable d'apparier une vidéo aux articles dans un maximum de cas, pour minimiser autant que possible l'intervention humaine, tout en maximisant le revenu.

Le principal verrou à surmonter tient au fait que cette collection interrogée par le système n'est pas exhaustive au regard de l'ensemble des sujets abordés par les articles traités. Cela signifie que les articles reçus n'ont pas tous de vidéos pertinentes en collection et certains n'ont donc aucune vidéo à intégrer en bas de page à l'issue du traitement. Or dans la perspective industrielle de ce projet, l'objectif est de maximiser l'intégration de vidéos en bas d'articles, afin d'optimiser le revenu. En effet, si aucune vidéo de contenu ne s'affiche dans la page, aucune publicité n'est jouée avant et aucun revenu n'est alors généré.

Il s'agit donc, pour le système mis en place, de trouver un compromis satisfaisant ces différentes exigences, loin d'être complémentaires.

Contexte scientifique

Les problématiques relatives au besoin industriel formulé par Mediabong s'articulent autour d'une question centrale : Qu'attend l'entreprise lorsqu'elle nous demande d'*apparier* une vidéo à un article ?

Cette notion d'*appariement* est floue parce qu'il est difficile de définir objectivement les fonctions qu'une vidéo doit remplir par rapport à un article. Une récente étude sur les nouveaux formats de presse en ligne détaille différents rôles que peuvent remplir les contenus audiovisuels dans les pages (Dagiral et al., 2010). Selon les auteurs, une vidéo peut être simplement illustrative et proposer en images l'exact contenu de ce que décrit l'article. Mais elle peut également tenir un rôle dénotatif ou connotatif : dans le premier cas, il s'agit par exemple d'une déclaration ou d'une manifestation relative au(x) fait(s) décrit(s) dans l'article ; dans le second cas, il s'agit plus d'un point de vue particulier, d'une interprétation possible du sujet de l'article. Dans ce dernier de cas, la vidéo associée à l'article peut d'ailleurs entrer en contradiction avec le contenu de celui-ci, bien qu'elle y

reste liée par le sujet abordé.

Si cette notion d'appariement se définit difficilement, c'est selon nous parce qu'elle repose sur le principe de *similarité*, constituant lui-même un mécanisme complexe. Considérer que deux objets sont similaires et donc susceptibles de former une paire, c'est d'abord définir les caractéristiques du type d'objet considéré puis définir des critères pour en comparer différentes occurrences.

Dans notre contexte d'appariement de contenus médiatiques, nous envisageons les textes comme des objets particuliers qu'il s'agit de comparer. Se posent alors des questions relatives à la façon de comparer ces objets textuels, ainsi qu'aux critères mobilisés pour rendre compte de leur degré de similarité. L'un de ces éléments est selon nous la similarité thématique de leurs contenus, c'est-à-dire le fait qu'un article et une vidéo relatent un même sujet principal. Prenons l'exemple **(1)** ci-dessous[3], la **Vidéo A** pourrait être appariée à l'article sur la base de leur similarité thématique, puisqu'il relate le même sujet. En revanche, ce même article ne pourrait former une paire pertinente avec la **Vidéo B**, qui décrit un sujet complètement différent.

(1)

Article	-	Les filles de Jacqueline Sauvage ont déposé une demande de grâce totale à l'Elysée.
Vidéo A	-	François Hollande gracie Jacqueline Sauvage : "Ce pouvoir ne devrait plus exister"
Vidéo B	-	Chemise déchirée à Air France : prison avec sursis pour trois ex-salariés

Partant de cette hypothèse, le premier enjeu est de modéliser automatiquement ces contenus de façon à ce qu'une machine puisse saisir le sujet qu'ils abordent. Pour y répondre, nous supposons que des méthodes liées au Traitement Automatique des Langues (TAL) pourraient être efficaces et permettre une représentation fine des textes.

C'est dans une perspective de recherche d'information (RI) que nous reconsidérons notre problématique. Dans notre contexte, nous considérons les articles comme les requêtes pour lesquels la vidéo à apparier constitue un document pertinent. Nous nous inspirons de méthodes proposées dans l'état de l'art pour deux étapes fondamentales du système d'appariement développé :

1. L'indexation des documents, *i.e.* la transformation d'un texte non structuré en une représentation structurée, rendant accessible sons sens par un système automatique.
2. La fonction mesurant le degré de similarité thématique entre différents documents, ou plutôt, entre un article et les vidéos de la collection interrogée.

3. Dans un souci de lisibilité, nous ne présentons ici que les titres des documents, mais ceux-ci sont bien accompagnés d'une description ainsi que de métadonnées (*cf.* Annexes A et B pour des exemples d'articles et de vidéos.)

Les approches proposées en RI ne sont toutefois pas suffisantes à nos yeux pour opérer l'appariement de contenus textuels médiatiques. Si la similarité thématique est un critère fondamental sur lequel repose l'appariement, la proximité temporelle des contenus est également primordiale lorsque l'on traite des données d'actualité. Un exemple éloquent de cette nécessité concerne les articles relatant un événement sportif particulier, dont les adversaires impliqués peuvent se rencontrer à différentes reprises au cours de leurs carrières. Prenons l'exemple du tournoi de tennis d'Open d'Australie. Il a vu s'affronter en finale Novak Djokovic et Andy Murray à la fois en 2011, 2013, 2015 et 2016. Un article paru en 2016, sur le match de cette année-ci, ne pourrait se satisfaire d'une vidéo datant de l'une des années précédentes. Il ne s'agirait pas du même événement, bien que les mêmes joueurs soient impliqués dans le même tournoi. La *fraîcheur* de la vidéo retournée est alors considérée dans nos travaux comme un facteur important dont dépend l'appariement. Les problématiques que soulève cette prise en compte temporelle relèvent de travaux ayant trait au domaine du *Topic Detection and Tracking* (TDT). Émergeant de la nécessité de traiter efficacement les flux continus d'actualités en ligne, ce domaine s'attache notamment à la comparaison de contenus médiatiques, à forte dimension temporelle.

Une fois les critères de comparaison établis, il s'agit ensuite de déterminer à partir de quel degré de similarité entre deux objets on peut considérer que la paire qu'ils forment est pertinente. En l'occurrence, lorsque l'on nous présente une paire article-vidéo, pour laquelle on a estimé la similarité thématique et la proximité temporelle, comment décider si l'appariement est pertinent ou non ?

La notion de *pertinence* est elle aussi complexe à définir, notamment dans le contexte de cette thèse, dont le but est de répondre à une ensemble de besoins divergents. La pertinence d'un appariement repose-t-elle uniquement sur les critères évoqués précédemment ? Nous savons que la collection vidéo n'est pas en mesure de proposer des contenus répondant à ces critères pour tous les articles traités. Mais parallèlement, les contraintes économiques du projet nécessitent d'intégrer une vidéo aux articles dans le maximum de cas. La réponse semble alors être que non. Si l'on souhaite associer une vidéo pour tous les articles ou presque, l'un des deux critères devrait être favorisé au détriment de l'autre, qui pourrait même ne pas être considéré du tout. Mais comment déterminer la façon de considérer ces différents critères dans la tâche d'appariement qui nous incombe ?

C'est en s'appuyant sur des méthodes relatives à la recommandation de contenus que nous avons tenté de répondre à ces problématiques. Les systèmes automatiques développés dans ce cadre cherchent à modéliser les préférences de leurs utilisateurs en se basant sur leur comportement vis-à-vis des résultats qui leur sont proposés. Les processus impliqués sont donc itératifs et le profil utilisateur évolue à chacun des jugements qu'il fournit.

Inspirés par ces méthodes, nous avons tenté de modéliser les attentes des utilisateurs du système d'appariement développé. N'ayant pas accès aux données des utilisateurs fi-

naux du système[4], nous travaillons sur celles fournies par un ensemble de salariés de MEDIABONG, en charge de la vérification des sorties du système. À partir des jugements de pertinence qu'ils attribuent à une série d'appariements automatiques, nous avons pu apprendre un modèle approchant leur comportement. Il s'agit d'une généralisation faite à partir d'exemples annotés manuellement et ce sont ici des méthodes d'apprentissage automatiques qui ont été implémentées.

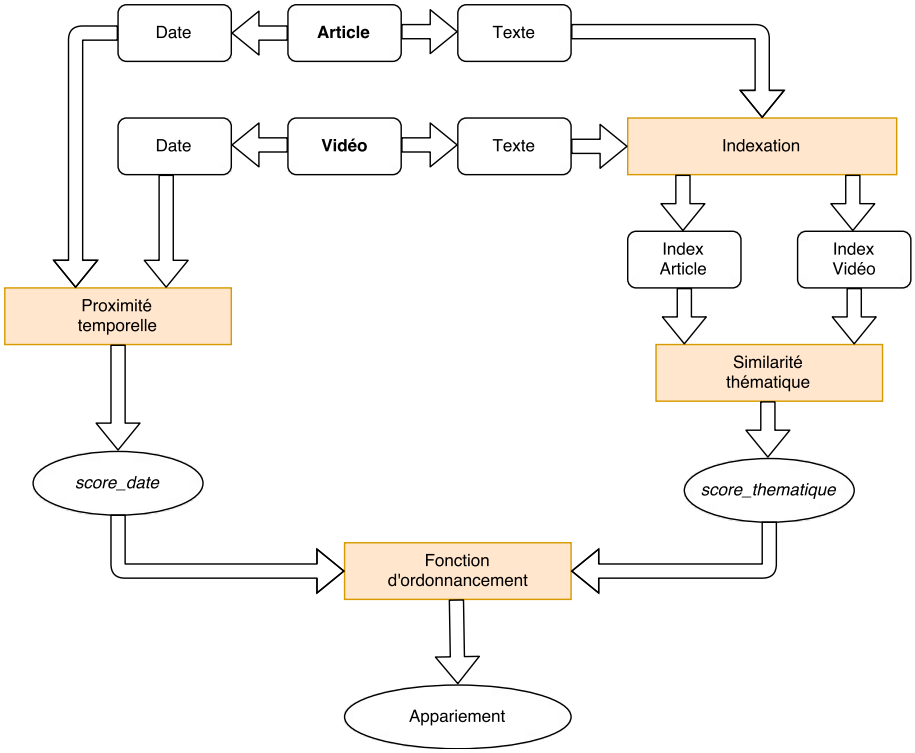

FIGURE 2 – Architecture générale du système d'appariement article-vidéo

En résumé, cette thèse met en œuvre des méthodes du TAL et de la RI pour développer un système d'appariement de contenus textuels médiatiques, en tenant compte des spécificités de ces contenus et des attentes de MEDIABONG. Ce système, illustré en Figure 2, distingue différentes étapes qui se complètent.

Pour clore cette section, notons que ces travaux de thèse ont donné lieu à une publication dans les actes de la 23ème édition de la conférence TALN :

— Désoyer, A., Battistelli, D., & Minel, J. L. (2016). Appariement d'articles en ligne et de vidéos : stratégies de sélection et méthodes d'évaluation. In *Actes de la conférence conjointe JEP-TALN-RECITAL 2016*, 2, 342-348.

De précédents travaux auxquels nous avons contribués ont par ailleurs participer à l'acquisition de certaines connaissances exploitées dans cette thèse, au sujet notamment des méthodes d'apprentissage automatique :

4. *i.e.* les internautes consultant un article en ligne auquel le système a associé une vidéo.

— Désoyer, A., Landragin, F., Tellier, I., Lefeuvre, A., & Antoine, J. Y. (2015). Les coréférences à l'oral : une expérience d'apprentissage automatique sur le corpus ANCOR. *Traitement Automatique des Langues*, 55(2), 97-121.

— Désoyer, A., Landragin, F., & Tellier, I. (2015). Apprentissage automatique d'un modèle de résolution de la coréférence à partir de données orales transcrites du français : le système CROC. In *Vingt-deuxième Conférence sur le Traitement Automatique des Langues Naturelles* (TALN'2015), 439-445.

— Désoyer, A., Landragin, F., Tellier, I., Lefeuvre, A., Antoine, J. Y., & Dinarelli, M. (2016). Coreference Resolution for French Oral Data : Machine Learning Experiments with ANCOR. In *17th International Conference on Intelligent Text Processing and Computational Linguistics* (CICLing'2016).

Organisation de la thèse

Ce mémoire, aboutissement des travaux menés tout au long de cette thèse, s'articule en trois parties. Nous avons tenté, dans chacune d'elles, de mettre en perspective les problématiques industrielles à l'origine de ce projet avec les travaux de recherche dans les domaines dont elles relèvent.

Nous débutons dans la Partie I par une discussion sur la notion d'appariement, au cœur de nos problématiques. Nous revenons sur les facteurs qui participent selon nous à l'acte d'appariement, celui de lier deux objets entre eux. Nous discutons d'abord du principe de similarité que nous considérons comme le pilier de l'appariement : deux objets sont en effet jugés similaires par un individu s'ils partagent un certain nombre de caractéristiques. Nous nous interrogeons sur la nature des traits mobilisés pour une telle opération et revenons ensuite particulièrement sur la comparaison d'objets textuels. La notion de pertinence est ensuite débattue. Régulièrement convoquée pour rendre compte de la *qualité* d'un résultat de système automatique, elle reste compliquée à saisir dans le contexte des sciences de l'information où aucun consensus ne semble avoir été trouvé quant à sa définition.

La Partie II est consacrée à une présentation de l'état de l'art des domaines de recherche dont relèvent nos travaux. Trois chapitres la composent.

Nous présentons au Chapitre 1 une revue des méthodes de résolution de la tâche de RI. Vaste domaine d'intérêt depuis les années 1950, la RI se donne pour objectif de retrouver des données non structurées dans une collection documentaire en réponse à un besoin d'information formulé par un utilisateur. L'automatisation de ce processus distingue généralement deux fonctions. La première est une fonction de modélisation des contenus dont le but est de transformer le texte initial en une représentation structurée

qu'une machine peut appréhender. La seconde est une fonction de comparaison qui cherche à saisir le degré de similarité entre la requête formulée par l'utilisateur et les documents de la collection susceptibles d'y répondre. Le chapitre se clôt par une présentation des méthodes d'évaluation de ces systèmes de recherche d'information (SRI), étape cruciale et non moins difficile à cerner, largement débattue au sein de la communauté.

Le Chapitre 2 poursuit cet état de l'art par la présentation de celui du domaine du *Topic Detection and Tracking*. S'intéressant aux données particulières que sont les informations d'actualité, les recherches investies dans ce domaine se décomposent en différentes tâches spécifiques sur lesquelles nous revenons. Parmi elles, la méta-tâche de *Story Link Detection* nous intéresse particulièrement. Elle soulève des problématiques relatives à la modélisation de contenus d'actualité, considérant notamment la dimension temporelle que nous souhaitons également prendre en compte dans nos travaux. Nous concluons ce chapitre par la présentation des campagnes destinées à l'évaluation des systèmes développés dans le domaine, ainsi que des corpus d'actualités exploités dans ce but.

Cette première partie se termine par le Chapitre 3, présentant succinctement les propositions académiques pour l'automatisation de la tâche de recommandation. Discipline plus récente que les précédentes, née dans les années 1990, la recommandation de contenus a émergé de la nécessité de considérer des paramètres individuels pour filtrer les informations de plus en plus massives sur le Web. Nous revenons sur différentes approches proposées pour modéliser les préférences des utilisateurs, puis sur les propositions de fonction de filtrage d'information. Le chapitre s'achève par une brève revue des méthodes d'évaluation des systèmes de recommandation.

La Partie III, composée de trois chapitres, présente finalement les contributions de cette thèse.

Nous décrivons au Chapitre 4 le système d'appariement d'articles et de vidéos développé en réponse au besoin industriel, baptisé SEMIABONG. Nous y exposons tout d'abord la façon dont les contenus des articles et vidéos sont représentés, en détaillant les étapes de la chaîne de traitement qui leur est appliquée. La sortie de cette chaîne, constituant la fonction d'indexation des contenus, distingue différents types d'unités et offre aux documents une représentation plus précise qu'un classique *sac de mots*. Les documents sont ensuite considérés dans SEMIABONG sous forme de vecteurs de termes, dont on saisit le degré de similarité via le calcul du cosinus de l'angle qu'ils forment. La valeur obtenue constitue ce que l'on appelle le *score_thématique* d'une vidéo pour un article. S'ajoute à celui-ci un *score_date*, mesurant la proximité temporelle entre les deux documents pour laquelle nous proposons une fonction empirique. Nous expliquons finalement la façon dont ces deux scores sont considérés dans l'ordonnancement final des résultats, grâce à l'exploitation des jugements de pertinence d'un échantillon d'utilisateurs du système.

Le Chapitre 5 revient sur la méthode d'évaluation de SEMIABONG, en débutant par les questions qu'elle a soulevées puis l'approche proposée pour y répondre. Dans notre cadre industriel, la performance *système* n'est en effet pas l'unique qualité à considérer pour rendre compte de son efficacité. Le besoin à satisfaire ne relève pas que d'un besoin d'information et le protocole d'évaluation standard en RI, le paradigme de Cranfield, s'est révélé insuffisant pour estimer les performances que nous cherchions à atteindre.

Malgré ses lacunes, ce protocole traditionnel permet de comparer les résultats de différents systèmes dans un cadre purement expérimental. Le Chapitre 6 apporte donc une contribution plus académique en proposant une série d'expérimentations menées dans un tel contexte. Nous y comparons différentes configurations de système, variant selon plusieurs paramètres, afin d'observer leurs résultats respectifs sur un même ensemble de test. Cet ensemble est construit suivant les règles classiques de l'évaluation en RI, en mobilisant des méthodes de *pooling* en amont de l'annotation manuelle des résultats de référence. Il constitue ainsi une ressource ouverte, exploitable par les membres de la communauté s'intéressant à l'association de contenus textuels médiatiques.

La Conclusion de ce manuscrit rappelle et synthétise l'état de l'art présenté ainsi que les contributions apportées. Sont ensuite exposées les perspectives de recherche que ces travaux nous ont permis d'envisager.

Première partie

Discussion liminaire : la notion d'appariement

Dans cette discussion liminaire, nous exposons le fruit des réflexions que ces travaux de thèse ont permis d'initier. C'est en effet après le développement du système d'appariement de contenus et de l'étude de l'état de l'art des domaines de recherche affiliés que les questions relatives aux notions présentées ici se sont dégagées. Ces notions de *similarité*, d'*appariement* et de *pertinence*, dont les définitions nous paraissaient assez intuitives en début de thèse, se sont révélées complexes au fur et à mesure de l'avancement de nos travaux. Essentielles selon nous à la compréhension du projet et aux problématiques qu'il soulève, nous avons souhaité présenter ces notions ici, dès le début de ce mémoire, bien que les questionnements à leur sujet ne soient intervenus qu'*a posteriori* de la réalisation du système automatique.

Nous nous interrogeons donc dans cette première partie sur la notion d'*appariement*, au cœur de ces travaux de thèse. Elle est définie dans (Littré et al., 1869) comme l'*action d'apparier, d'unir par couple, d'assortir par paire*. Une paire y est par ailleurs définie comme *deux choses de même espèce qui vont ensemble*.

Découlent de ces définitions élémentaires une foule de questions que nous souhaitons approfondir ici. On se demande notamment comment s'établissent les critères décrétant que deux choses *vont ensemble* ? Sont-ils totalement subjectifs et personnels, ou est-il possible d'en objectiver certains, d'en dégager des universaux ?

Dans une tâche d'appariement, la première question qui se pose selon nous est celle des critères retenus pour sélectionner les objets à unir par paire. En d'autres termes, comment considérer que deux objets parmi un ensemble sont considérés comme similaires ? De notre point de vue, intervient ensuite une seconde question relative aux critères d'appréciation d'une paire d'objets. Sur quelle qualité de la paire se baser pour considérer qu'elle est *correcte*, c'est-à-dire qu'il s'agit bien d'une paire ? La notion de *pertinence*, largement exploitée dans le domaine de la RI, est-elle suffisante pour répondre à cette question ?

Cette discussion commence donc par présenter en Section 1 le principe de *similarité*, sur lequel repose la tâche d'*appariement*. Après en avoir évoqué certaines généralités d'un point de vue psychologique et cognitif, nous nous intéresserons particulièrement à la similarité d'un point de vue linguistique. Nous reviendrons notamment sur les approches proposées dans la littérature pour mesurer la similarité entre différents objets textuels puisque cette tâche est à l'origine notre problématique.

Nous discuterons ensuite en Section 2 des critères d'appréciation mobilisés par l'humain pour juger de la qualité d'une information délivrée par un système automatique. Nous baserons cette discussion sur des travaux en sciences de l'information, qui s'intéressent particulièrement aux questions relatives à l'évaluation de l'information, et reviendrons particulièrement sur le critère de *pertinence*.

Pour clore cette discussion, nous mettrons en perspective en Section 3 les notions

abordées en amont avec les données de notre projet. Nous présenterons quelques exemples d'appariement d'article et de vidéo proposés par le système développé, ainsi que les jugements d'un échantillon d'utilisateurs à leur égard.

Principe de similarité

Lorsque l'on souhaite organiser un ensemble d'objets, en regrouper certains parmi eux, on se base sur certaines de leurs caractéristiques communes. Mesurer la distance entre différents objets, à partir de l'observation de ces traits partagés, relève de ce que l'on appelle la *similarité*.

Nous revenons ici sur cette notion, par une présentation succincte des questions générales qu'elle soulève du point de vue de la représentation mentale de la tâche qu'elle incarne. La comparaison d'objets textuels, qui constituent l'objet d'étude de cette thèse, est ensuite discutée, autour de notions relatives à la similarité linguistique.

Généralités

Les interrogations sur la notion de similarité relèvent en premier lieu de recherches en psychologie, notamment sur la cognition. Comment l'humain saisit-il la similarité entre différents objets d'une collection ? Sur quels critères se base-t-il ? Tous les individus mobilisent-ils les mêmes critères pour ce faire ?

Le psychologue Amos Tversky estime que la similarité joue un *"rôle fondamental dans les théories de la connaissance et du comportement. C'est un principe organisateur par lequel les individus classent les objets, forment des concepts et généralisent."* (Tversky, 1977). La similarité est en effet un principe qui régit nos quotidiens, sans d'ailleurs que l'on en soit toujours bien conscient. Le fait est qu'il est souvent plus simple de retrouver un objet dans un ensemble organisé que lorsque règne l'anarchie au milieu de laquelle on peut se perdre. C'est pour cette raison que l'homme à tendance à ranger, trier, organiser les choses.

Illustrons ce propos par un exemple concret. Lorsque l'on décide de ranger sa bibliothèque, on opte régulièrement pour une classification qui nous permettra de retrouver efficacement ce que l'on souhaite lorsqu'on en aura besoin. Ainsi, on choisit régulièrement le classement par ordre alphabétique d'auteurs, qui reste un critère simple et efficace de tri. Mais certaines organisations peuvent être plus sophistiquées, et distinguer en plus les livres par genre, par origine de l'auteur, par langue si la collection est multilingue, *etc.*

Toute tâche de catégorisation nécessite en amont de définir les traits sur lesquels vont être regroupés les objets. C'est ensuite que la notion de similarité intervient, puisqu'une

fois les traits définis, on cherche à mettre ensemble ceux partageant les mêmes valeurs sur ces traits.

La complexité de la notion de similarité est due à la complexité des objets qu'elle cherche à comparer, ceux-ci étant composés de multiples dimensions. La Figure 3 illustre cette complexité par un exemple rudimentaire. Elle présente divers objets qui forment un ensemble, dont certains présentent des caractéristiques communes susceptibles de servir de critères pour distinguer différents groupes d'objets.

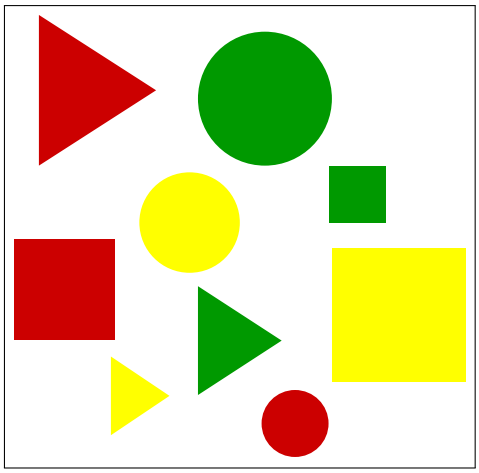

FIGURE 3 – Un ensemble d'objets à classer

Si l'on demande à un individu de construire, à partir de cet ensemble, trois groupes d'objets, quel sera le résultat ? Il a en effet le choix de regrouper les objets selon leur forme, leur couleur ou leur taille. Notons que cette dernière variable est relative, et ne peut être mobilisée que si l'on considère l'ensemble des objets : on ne peut dire en effet qu'un objet est petit en soi, il l'est par rapport à un autre qui est plus grand. Les deux autres dimensions, forme et couleur, sont en revanche des caractéristiques intrinsèques aux objets.

Le résultat de cette tâche de regroupement pourra donc être différent d'un individu à l'autre, si l'on ne leur fournit pas plus de précisions dans l'énoncé. En revanche, si l'on formule cet énoncé en ajoutant une information sur la dimension à considérer, en demandant par exemple de créer trois groupes d'objets sur la base de leurs formes communes, les résultats devraient être plus homogènes.

Dans une tâche légèrement différente, lorsque l'on demande à un individu de juger une paire d'objets créée manuellement ou automatiquement, voire aléatoirement, sur quel(s) critère(s) va-t-il se baser pour l'accomplir ? De nouveau dans cette tâche-ci, il est préférable de fournir aux individus engagés les règles à appliquer pour juger une paire si l'on souhaite obtenir des jugements les plus objectifs possibles. Sans cela, on prend le risque de se retrouver avec des jugements différents sur une même paire.

Prenons par exemple les deux objets présentés en Figure 4. En demandant à un jury d'évaluateurs de juger de la pertinence de cet appariement, sur une échelle binaire distinguant les pertinents des non pertinents, on pourra observer différents résultats en fonction des juges. Certains pourront considérer l'appariement pertinent, du fait que les objets qui la composent ont la même forme. En revanche, un juge s'intéressant particulièrement à l'appariement des couleurs pourra juger cette paire non pertinente.

En signalant aux juges, en amont de l'évaluation, que l'objectif est d'apparier des objets sur la base de leur forme commune, tous devraient s'accorder pour considérer que cet appariement est pertinent. L'emploi du conditionnel ici tient au fait que, même lorsqu'un protocole d'évaluation clair est fourni aux juges, il est toujours envisageable que certains puissent se tromper sur certains des cas présentés. Sciemment ou non, tout juge humain peut en effet fournir une réponse allant à l'encontre de ce qui est attendu au regard du protocole.

En résumé, que ce soit pour une tâche de création ou d'appréciation de paire d'objets, il semble nécessaire d'établir des règles sur lesquelles les individus à qui l'on demande leurs avis puissent se reposer.

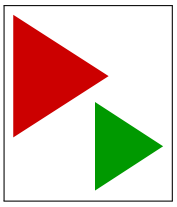

FIGURE 4 – Une paire d'objets à juger

Si l'on veut modéliser plus formellement cette problématique de la similarité, on peut envisager un espace dans lequel chacune des dimensions représente une caractéristique des objets considérés (*cf.* Figure 5). Chaque individu se construit mentalement son propre espace, y projette les objets qu'il cherche à comparer, puis juge de leur degré de similarité en fonction de la distance qui sépare ces objets dans cet espace (THIBAUT, 1997). Or tous les individus ne construisent pas forcément le même espace. Ils peuvent ne pas considérer les mêmes dimensions, ou s'il le font, ne pas accorder la même importance à chacune d'elles. Intervient alors ici la notion de pondération des dimensions, et s'ensuit une interrogation sur la notion de saillance de certaines caractéristiques des objets. Tous les traits composant un objet participent-ils également à sa représentation dans l'esprit d'un individu ? Ou existe-t-il une hiérarchie dans les propriétés le décrivant ?

Pour les objets physiques, tels qu'illustrés dans les exemples précédents, il est communément admis que la forme soit la propriété saillante, devant la couleur ou la taille. Cela s'explique par le fait qu'elle constitue la propriété fondamentale de l'objet d'après (LANDRAGIN, 2011), et s'impose donc devant les autres. L'auteur présente en fait une

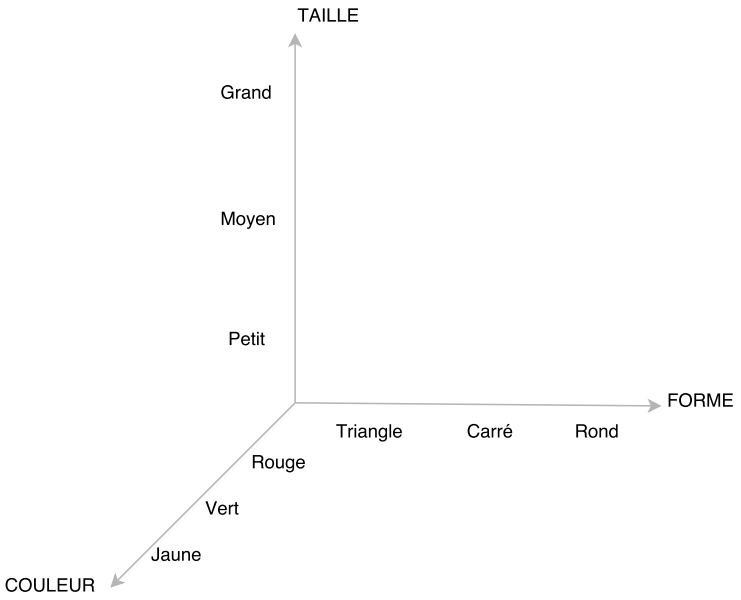

FIGURE 5 – Modélisation des dimensions élémentaires d'un objet physique

analogie entre saillance visuelle des objets physiques et saillance linguistique des objets textuels. Ces derniers sont ceux qui nous intéressent particulièrement dans le cadre de cette thèse.

Dans la suite de cette discussion, nous nous focalisons sur la comparaison d'objets linguistiques, et plus spécifiquement de textes. Nous nous interrogeons sur les traits définissant ces objets particuliers, sur la façon d'opérer une comparaison entre eux, ainsi que sur les critères d'évaluation de la similarité entre textes.

Similarité textuelle

Un texte est un objet particulier composés d'éléments linguistiques à différents degrés. Communément dans le domaine, on distingue les niveaux morphologique, lexical, syntaxique et sémantique pour décrire ces objets. Bien que d'autres niveaux intermédiaires puissent être considérés, ces derniers caractérisent les composants élémentaires du langage.

Lorsque l'on se donne pour tâche de mesurer la similarité entre deux objets linguistiques, on doit, de la même façon que pour les objets physiques, s'interroger sur les critères à considérer pour opérer une telle comparaison. En 2013, une étude sur les différents types d'approche pour mesurer automatiquement cette similarité est proposée (GOMAA et al., 2013). Depuis l'explosion du Web à la fin des années 1980, les propositions pour automatiser cette tâche sont devenues légion. Mesurer la similarité entre des textes, ou des portions de textes, est en effet devenue la fonction élémentaire de bon nombre d'applications liées à la recherche d'information, la classification ou le regroupement de contenus, la détection de thématiques, le résumé automatique, *etc.*

Pour répondre aux problématiques qu'impliquent ce défi, les chercheurs se sont mobilisés autour des questions relatives à la représentation des contenus et aux fonctions mesurant la similarité entre ceux-ci. En s'inspirant des représentations mentales que se font les individus des textes, ils ont proposé des modélisations numériques approchant ces représentations humaines. C'est dans ce cadre que nous situons cette discussion, en nous intéressant à la représentation d'un texte d'un point de vue informatique, et aux différents indices convoqués pour optimiser sa précision, *i.e.* sa capacité à rendre compte compte du sens initial du texte.

Les auteurs de (Gomaa et al., 2013) distinguent trois grandes approches pour mesurer la similarité textuelle. Toutes ces méthodes se distinguent par le niveau des informations considérées dans la représentation des textes : les mots ; les corpus ; les connaissances.

Ainsi, les premières méthodes proposées se basent sur la chaîne de caractères. D'un pont de vue linguistique, ces approches se situent donc à un niveau de représentation lexical. L'idée générale est de segmenter un texte en mots et d'en constituer un ensemble pour représenter celui-ci. La principale faiblesse de ces représentations est qu'elles ne conservent pas l'ordre initial des termes or, la linéarité est une des propriétés fondamentales du langage (Cotte, 1999) et sa perte entraîne avec elle celle de certaines informations, dont celles relatives aux niveaux de la syntaxe et de la sémantique.

C'est pour pallier ce biais-là que l'on a commencé à s'intéresser aux expressions multimots (EMM), considérant alors des suites de termes, souvent contiguës, dans les textes. Il s'agit d' « *unités linguistiques complexes qui contiennent un certain degré de non-compositionalité lexicale, syntaxique, sémantique et/ou pragmatique.* »(Constant, 2012). Elles recouvrent différents types de sous-unités, telles que les expressions figées (***e.g.*** *avoir du pain sur la planche, moulin à paroles*) ; les entités nommées (***e.g.*** *Ferdinand de Saussure, Notre Dame de la Garde*) ; les constructions à verbes support (***e.g.*** *faire confiance à, avoir conscience de*) ou encore les collocations (***e.g.*** *taxe d'habitation, fruit rouge*). Même si elles ne permettent pas de conserver la linéarité initiale des textes, elles aident malgré tout à saisir un niveau intermédiaire entre lexique et syntaxe, pour saisir certaines informations pertinentes. Ces EMM sont plus précises que des termes simples (au sens monolexicaux) et réfèrent régulièrement à des entités extra-linguistiques uniques, d'où la nécessité de les identifier en tant que telles.

Parmi ces expressions, les syntagmes, notamment nominaux, et les noms propres sont jugés dans l'état de l'art comme particulièrement informatifs, c'est-à-dire particulièrement utile pour réduire la complexité des textes (M. Gross et Senellart, 1998). Par ailleurs, les noms propres sont très présents dans les articles de presse et représentent plus de 10% des termes que l'on peut y recenser (Friburger et al., 2002). Dans ces contenus particuliers, on observe que la mention de noms de personne aide le lecteur à circonscrire l'événement qui y est relaté.

Pour repérer et extraire automatiquement ces syntagmes nominaux (SN) et entités nommées (EN), la plupart des travaux se basent sur l'étude des corpus et non plus sur les textes indépendamment les uns des autres. C'est de la comparaison et du regroupement d'informations issues de la masse de données de ces différents textes qu'émergent ces composants. L'une des méthodes les plus répandues pour identifier ces unités est l'analyse des co-occurrences de termes via des mesures telles que le χ^2, l'information mutuelle ou le gain d'information (Zheng et al., 2003) qui permettent de repérer des séquences de mots régulièrement conjoints en corpus. Ces segments répétés s'avèrent sémantiquement plus riches que les mots qui les composent pris indépendamment (Salem, 1986).

Le dernier niveau de représentation des contenus décrit dans (Gomaa et al., 2013) fait appel à des réseaux sémantiques tels que WordNet [5] pour l'Anglais (Miller, 1995). L'interrogation de ce type de ressource permet d'obtenir une représentation sémantique des textes capable de faire émerger des concepts qui n'y figurent pas littéralement. Par exemple, un texte au sujet de *chats* peut être considéré comme proche d'un texte au sujet de *chiens*, puisque leurs représentations sémantiques mentionnent que ces deux concepts sont des sous-concepts de la catégorie plus générique *animaux de compagnie*. Plus récemment encore, les efforts se concentrent sur des méthodes d'apprentissage automatique à base de réseaux de neurones qui permettent de déduire des corpus des relations sémantiques entre des termes. il s'agit généralement de représenter les termes sous forme de vecteurs de traits et de les projeter dans un espace où il devient possible de de les combiner en opérations telles que :

$$vecteur(Roi) - vecteur(Homme) + vecteur(Femme) = vecteur(Reine)$$

Il n'est alors plus nécessaire d'interroger des ressources externes, particulièrement coûteuses à mettre en place. Ces méthodes, connues sous le nom de *word embedding* (ou plongement lexical, en français) sont de plus en plus implémentées dans les travaux de modélisation de la langue (Mikolov et al., 2013 ; Levy et al., 2014).

Au regard de ce rapide tour d'horizon, il apparaît que c'est au niveau sémantique que se joue l'essentiel des efforts en modélisation des contenus. Dans l'objectif de mesurer la similarité entre deux textes, il est en effet nécessaire de saisir en amont le *sens* véhiculé par chacun d'eux. La difficulté réside entre autres dans le fait que le sens n'est pas compositionnel, *i.e.* pas déductible de la somme des sens de chacun de ses composants.

Dans ses travaux, l'auteur de (Landragin, 2004 ; Landragin, 2012) décrit d'ailleurs les propriétés sémantiques comme des critères remarquables sous-tendant la saillance linguistique. Il énumère un certain nombre de caractéristiques permettant aux constituants d'être considérés comme saillants dans les textes. Ainsi, le nom propre apparaît en tête

5. https://wordnet.princeton.edu/

de liste, suivi du trait animé (*vs.* inanimé) et enfin du rôle d'agent (*vs.* patient). Les exemples **(1)**, **(2)** et **(3)** ci-dessous illustrent, dans l'ordre de citation, ces processus. Tous les exemples présentés par la suite sont extraits du corpus de données de MEDIABONG. Il s'agit de titres de différents articles reçus de leurs partenaires diffuseurs.

(1)
Les filles de **Jacqueline Sauvage** ont déposé une demande de grâce totale à l'Elysée.

(2)
Le pape, en visite à Prato, plaide pour un travail digne.

(3)
Trump dit à Xi qu'il respectera le principe d'"une seule Chine".

Des indices de saillance relevant d'autres niveaux linguistiques sont par ailleurs recensés par l'auteur. Au niveau syntaxique d'abord, la construction clivée **(4)**, le détachement en tête de phrase **(5)** détaillé dans (CHAROLLES, 2003) ou au contraire le rejet en fin de phrase **(6)** participent à mettre en relief certaines unités textuelles. Ces constructions particulières offrent un cadre physique à l'information que l'auteur souhaite souligner.

(4)
Dans le final en faux plat montant, *c'est* **le Belge Greg Van Avermaet** *qui* s'est montré le plus solide.

(5)
À Alep, les civils doivent aussi faire face à la soif, à la faim et au froid.

(6)
Montpellier : une journaliste frappée par des lycéens **lors d'une manifestation pro-Théo**.

(7)
Galileo : plusieurs horloges atomiques en panne.

(8)
5 idées reçues sur la **dépression** : Idée reçue #4 : La **dépression** est un problème de riches

(9)
Le secrétaire américain à la Défense rencontre des troupes anti-EI en Jordanie.

Peuvent s'ajouter à ces critères des indices typographiques, telle l'insertion de deux-points mettant en exergue une unité particulière **(7)**. La fréquence d'occurrence de certaines unités peut également être facteur de saillance. La raison sous-jacente est que si l'auteur fait le choix d'insister sur un point en le répétant, c'est qu'il souhaite que le lecteur y prête particulièrement attention **(8)**. Au niveau grammatical, l'auteur observe que

la position sujet est privilégiée à celle d'objet direct, elle-même privilégiée à celle d'objet indirect **(9)**.

Les auteurs de (Hatzivassiloglou, Klavans et al., 1999) présentent une méthode pour mesurer la distance sémantique entre deux textes courts relatant des événements d'actualité. Dans le cadre de leur étude, ils considèrent que deux contenus sont similaires s'ils portent le même intérêt à un même concept, acteur, objet ou action. Ce même acteur ou concept doit par ailleurs accomplir ou être soumis à la même action, ou être le sujet de la même description. Les auteurs illustrent cette définition par l'exemple présenté ci-dessous en **(10)** : selon eux, les unités **(10a)** et **(10b)** sont similaires puisqu'elles relatent le même événement (la perte de contact) et se focalisent sur le participant principal, l'hélicoptère. En revanche, **(10c)** n'est similaire ni à **(10a)** ni à **(10b)** puisqu'elle se focalise sur l'urgence d'atterrir due à la perte de contact, plutôt qu'à la perte de contact elle-même.

(10)

a - An OH-58 helicopter, carrying a crew of two, was on a routine training orientation when contact was lost at about 11 :30 a.m. Saturday (9 :30 p.m. EST Friday).

b - "There were two people on board," said Bacon. "We lost radar contact with the helicopter about 9 :15 EST (0215 GMT)."

c - An OH-58 U.S. military scout helicopter made an emergency landing in North Korea at about 9.15 p.m. EST Friday(0215 GMT Saturday), the Defense Department said.

En appliquant cette définition à notre corpus, les contenus des exemples **(11a)** et **(11b)** seraient considérés comme similaires, tandis que les paires formées par les exemples **(11a,11c)** et **(11b,11c)** seraient considérées comme dissimilaires.

(11)

a - Usain Bolt s'est entraîné avec une équipe de DH de la Côte d'Azur

b - Usain Bolt joue au foot avec les amateurs du JS St-Jean Beaulieu

c - "I am Bolt" le documentaire sur Usain Bolt

Nous détaillerons plus loin en Section Section 3 cette notion de similarité appliquée aux données de notre corpus. L'observation de jugements utilisateur sur des appariements automatiques entre articles et vidéos produits par notre système nous a en effet permis de mettre en perspective les conclusions d'études présentées ici avec nos propres données.

Avant cela, nous souhaitons poursuivre la discussion autour des critères considérés pour apprécier une paire d'objets, particulièrement d'objets textuels. Nous suggérions plus haut le fait qu'évaluer la similarité entre deux objets correspondait à juger du niveau de pertinence de la paire qu'ils forment. Or, que sait-on réellement de cette notion de *pertinence* ? Comment se définit-elle ? Et, par ailleurs, est-ce la seule qualité mobilisée pour juger du degré de similarité entre deux objets ?

Critère de pertinence

Nous nous intéressons dans cette section à la notion de pertinence dans le contexte des sciences de l'information, et plus spécifiquement dans celui de la recherche d'information. Nous revenons sur sa définition, ses points forts et ses faiblesses.

Nombre d'études au sujet des critères d'appréciation des résultats ont été menées dans un contexte de RI. Les problématiques qu'elles abordent cherchent à comprendre comment les utilisateurs d'un système de recherche d'information (SRI) jugent de la qualité des documents fournis en réponse à la requête formulée en entrée. Si la pertinence est largement convoquée pour répondre à ces questionnements, elle ne fait pas consensus quant à sa définition ni à son objectivité. En effet, l'appréciation d'un résultat semble dépendre de l'objectif de l'utilisateur et du contexte de RI dans lequel s'inscrit sa démarche. En fonction de ceux-ci, il s'avère que la similarité entre les contenus n'est pas toujours suffisante pour expliquer les jugements et les choix des utilisateurs.

Comme la majorité des productions scientifiques, celles ayant trait à la notion de pertinence en sciences de l'information sont essentiellement rédigées en anglais. Or, lorsque l'on évoque la notion de pertinence en français, l'anglais la traduit généralement par *relevance*, mais peut aussi y référer sous le terme de *pertinence* ou encore d'*aboutness*.

Cette variété de dénominations est très révélatrice de la difficulté à définir ce concept de façon consensuelle. Il semble en premier lieu que sa définition dépende du point de vue duquel on l'envisage. Si l'on se place dans une perspective psychologique, du point de vue de l'utilisateur, ou à l'inverse dans une perspective logique, du point de vue du système, l'approche diffère dès le départ (Cooper, 1971).

L'auteur de (Saracevic, 1975), chercheur en sciences de l'information, présente la distinction entre les deux termes *pertinence* et *relevance* en parallèle de celle opposant les concepts de besoin d'information et de requête formulée. Un SRI peut renvoyer des résultats *relevant*, *i.e* qui répondent correctement à la requête soumise, mais pas nécessairement *pertinent*, *i.e.* qui ne répondent pas au besoin d'information initialement formé dans l'esprit de l'utilisateur.

Le besoin d'information se définit comme un état psychologique associé à l'incertitude et au désir de savoir ce qui est inconnu. (Cooper, 1971) écrit à son sujet que bien qu'il ne soit pas observable, car abstrait, il a une existence propre dans l'esprit humain et il est donc utile en cela de le nommer.

Il s'avère assez délicat de formuler une question exprimant ce que l'on se demande effectivement. Lorsque l'on s'interroge sur un sujet dont on ignore tout, il est particulièrement difficile de formuler à ce sujet une quelconque question. Comment savoir ce que l'on veut savoir puisque, logiquement, on l'ignore avant de le savoir ?

Il existe en fait différentes démarches menant à un processus de RI, c'est-à-dire différents types de besoin d'information. Il peut être très précis, et formulé clairement, comme lorsque l'on se demande *"En quelle année Christophe Collomb a-t-il découvert l'Amérique ?"* ou encore *"Berlin est-elle la capitale de l'Allemagne ?"*. Dans ces cas particuliers, la réponse attendue est unique et sans équivoque. En revanche, lorsque l'on s'intéresse à un sujet qui nous est inconnu, sur lequel on souhaite en apprendre plus, aucune réponse précise n'est attendue. Si l'on soumet à un moteur de recherche une requête du type *"Tout connaître de la physique nucléaire"*, les résultats risquent d'être nombreux et variés. Le jugement n'en sera que plus difficile, d'autant que si l'on ne sait pas bien ce que l'on cherche, on ne saura pas forcément apprécier ce que l'on nous propose.

La notion de pertinence situationnelle est proposée dans (Wilson, 1973), c'est-à-dire le fait qu'une information ne peut être pertinente en soi, mais peut l'être pour un certain individu dans un certain contexte. Chaque individu ayant sa propre vision du monde, et son propre lot de connaissances sur certains sujets, il apparaît en effet qu'une même information délivrée à différents individus n'aura pas sur chacun d'eux le même impact.

Dans le même esprit, la pertinence est décrite dans (Saracevic, 1975) comme la relation entre la connaissance d'un individu sur un sujet donné, ou disons son état de connaissance sur ce sujet à un moment donné, et la connaissance générale sur ce même sujet. La pertinence permet alors de mesurer le changement d'un état de connaissance d'un individu particulier. Un résultat de SRI sera jugé pertinent si l'état de connaissance de l'utilisateur concernant le sujet de sa requête a évolué après avoir consulté ce résultat.

La mesure de la pertinence est par conséquent hautement subjective, puisque dépendante des individus. On peut lire dans (Voorhees, 2000), reprenant les termes de (Schamber, 1994), que *"les jugements de pertinence sont connus pour varier d'un juge à l'autre, et d'un moment à l'autre pour un même juge"*[6]. Cette forte subjectivité, tant inter- qu'intra-personnelle, fait de la pertinence un critère instable pour l'évaluation des performances d'un SRI.

C'est pour passer outre ce biais qu'est définie dans (Cooper, 1971) la *pertinence logique*, qui permet à l'auteur de distinguer la pertinence de l'utilité d'une réponse à un besoin d'information. Dans son paradigme, il postule que la pertinence relève de l'*aboutness*[7], c'est-à-dire le fait que le sujet d'une réponse soit le même que celui de la requête. La pertinence relève en cela d'implications logiques, basées sur la similarité linguistique unissant la requête et les documents pertinents. L'utilité quant à elle est bien plus large, et relève d'aspects variés et subjectifs d'une réponse, telle que la qualité de la source, sa crédibilité ou encore son actualité. À partir de cette distinction fondamentale, l'auteur postule qu'un

6. *"Relevance judgments are known to differ across judges and for the same judge at different times."*
7. Cooper parle aussi de *topicality* pour décrire la pertinence logique. Il s'agit bien du fait, pour un résultat, de détailler le même sujet que celui de la requête.

résultat de SRI peut être utile sans être pertinent ou inversement pertinent sans être utile en réponse au besoin d'information initial.

Quel compromis trouver alors entre ces deux qualités ? Mieux vaut-il ne se consacrer qu'à l'utilité et négliger la pertinence logique ? Ou ne considérer que la pertinence logique qui, d'un point de vue algorithmique, est nettement plus *facile* à appréhender et ce malgré la complexité de la langue ?

Si les textes sont en effet des objets complexes, difficiles à modéliser, le comportement humain l'est encore d'avantage. La subjectivité dont leurs choix et jugements sont empreints freinent les espoirs d'une représentation générique et standardisée. Toutefois, la discussion présentée dans (Cooper, 1971) se conclue par une question de l'auteur se demandant si la pertinence logique ne constituerait pas un facteur, si ce n'est le facteur majeur, affectant l'utilité d'un résultat. Les deux ne semblent en effet pas complètement indépendants. Si l'on reprend l'exemple de l'utilisateur s'interrogeant sur le domaine de la physique nucléaire, on peut considérer qu'un document décrivant ce qu'est un noyau atomique est à la fois pertinent, puisque sémantiquement lié au sujet de la recherche, et utile, puisqu'il répond au moins en partie aux interrogations de l'utilisateur. Sa connaissance évolue sur le sujet en consultant ce résultat, en faisant un résultat pertinent selon (Saracevic, 1975).

On constate finalement que les différentes définitions proposées se recoupent entre elles. Toutes ne sont simplement pas appréhendées selon le même point de vue. Il apparaît que la pertinence mesure le degré d'information apportée à un utilisateur en réponse à un besoin traduit dans une requête. Or même si ce n'est pas le seul critère convoqué, la similarité des contenus de la requête et des documents retournés semble essentielle à l'estimation de leur pertinence.

Dans le cadre de notre étude, c'est une tâche d'appariement de contenus que l'on cherche à remplir automatiquement. Dans un tel contexte, les juges sont en charge d'évaluer des paires de textes médiatiques proposées par le système. Ils ne sont donc pas à l'origine de la démarche de RI, n'ont aucun besoin d'information à combler et ne formulent aucune requête. Le paradigme change donc par rapport à celui des études présentées ici, puisqu'il n'y a pas d'évolution d'état de connaissance d'un instant t à un instant $t+1$. Toutefois, la notion de pertinence peut être mobilisée par les utilisateurs pour juger la paire qui leur est proposée. C'est pourquoi il nous a semblé opportun ici de présenter cette notion et l'ambiguïté qui la caractérise.

En outre, si notre problématique ainsi décrite s'apparente à une tâche de recommandation de contenus, dont nous présentons un état de l'art au Chapitre 3, elle s'en distingue par une caractéristique fondamentale, celle des utilisateurs impliqués dans le processus. En effet, dans la majorité des systèmes de recommandation, et notamment ceux de re-

commandation d'actualités, l'objectif est de proposer aux utilisateurs *réels* du système des contenus personnalisés, en fonction de ce que l'on connaît de leurs préférences. Le système déduit ces préférences des jugements de pertinence fournis par chacun des utilisateurs pour des articles qui leur ont été présentés antérieurement. Or, dans notre cadre d'étude, le système mis en place est destiné à une équipe de salariés de MEDIABONG, chargée de sélectionner la vidéo proposée pour un article avant l'intégration effective de celle-ci dans les pages des diffuseurs partenaires. C'est cette équipe [8] qui observe les recommandations proposées par le système et qui sélectionne finalement la vidéo à apparier à chaque article. Ce sont donc leurs jugements qui sont exploités dans nos expérimentations. Nous démontrons tout au long de cette thèse les difficultés que cette contrainte a posées, notamment pour l'étape d'évaluation des performances du système (Chapitres 5 et 6). Les irrégularités dans les choix d'appariement de cette équipe d'utilisateurs ont grandement compliqué nos recherches sur la compréhension et la modélisation de la pertinence d'une paire article-vidéo.

Nous aurions souhaité pouvoir accéder aux données des utilisateurs réels du système, *i.e.* les lecteurs des pages d'articles dans lesquelles les vidéos sont intégrées. Toutefois, pour des raisons à la fois pragmatique et technique, il ne nous a pas été possible de recueillir ces données. Pragmatique car il aurait été délicat de tester le système sur des utilisateurs qui sont en fait les clients des clients de MEDIABONG. En effet, MEDIABONG intègre des vidéos dans les pages de ses clients diffuseurs, pages dédiées à leurs visiteurs. Pour satisfaire les diffuseurs, il faut donc satisfaire leurs internautes en leur proposant des vidéos *pertinentes* par rapport aux articles, susceptibles de les intéresser. En envisageant de tester un système en l'évaluant à partir des retours de ces internautes, on prendrait le risque de proposer de mauvais appariements dans les sites qui pourraient décider, sur ce constat de mauvaise performance, de rompre le partenariat avec MEDIABONG. En filtrant les résultats en amont de l'intégration, grâce par l'équipe de MEDIABONG, on évite ce type d'écueil. La raison technique tient quant à elle au fait que nous n'avons pas accès au nombre de vues d'une vidéo par URL, information qu'il est nécessaire de connaître si l'on souhaite évaluer son appréciation par l'ensemble des internautes. Le calcul de la rémunération à verser à chaque producteur se fait sur la base du nombre de fois que ses vidéos ont été vues, et ce indépendamment de la page dans lesquelles elles sont lancées. Les diffuseurs sont également rémunérés sur le nombre d'impressions globales, toutes URLs confondues, auquel s'ajoute des paramètres dépendants du type de contrat passé avec MEDIABONG. Pour toutes ces raisons, c'est donc pour les salariés de MEDIABONG que le système développé dans cette thèse est conçu, et ce sont eux que nous désignons sous le

8. La taille de cette équipe a varié tout au long de la thèse, au gré des arrivées et départs de stagiaires et alternants impliqués dans la tâche le temps de leur contrat chez MEDIABONG. Seuls deux d'entre eux sont des salariés permanents qui ont participé à toutes les évaluations menées et présentées dans ce mémoire.

terme d'*utilisateurs*.

Cas particulier : l'appariement de contenus médiatiques

La tâche d'appariement de contenus médiatiques a été reconsidérée dans ces travaux comme une tâche particulière de RI. Un article soumis au système peut en effet être considéré comme une requête à laquelle la vidéo associée[9] peut être vue comme une réponse au besoin d'information concentré dans l'article.

Toutefois, cette notion de réponse suppose qu'il y ait eu question en amont. Et en effet, dans un cadre classique de RI, l'utilisateur part d'un besoin d'information pour formuler une requête à laquelle les résultats proposés doivent répondre. Dans notre contexte, il n'y a donc pas de besoin d'information clair, mais on peut en revanche considérer que le contenu d'un article constitue un sujet sur lequel on souhaite en savoir plus et en cela, la vidéo à apparier doit être relative à ce sujet-ci. Le système cherche alors à faire évoluer la connaissance de l'utilisateur sur le sujet que développe l'article, en lui apportant des informations relatives à ce sujet, *i.e.* illustratives et/ou complémentaires.

Nous revenons alors ici à la question de la similarité textuelle, en se demandant dans ce contexte ce qui fait qu'une vidéo est similaire à l'article. Par ailleurs, nous nous interrogeons sur les critères d'appréciation mobilisés par les utilisateurs pour juger de la qualité d'un appariement. La pertinence, au sens de (Cooper, 1971), décrit la qualité d'un résultat décrivant le même sujet que celui de la requête. Nous supposons ici que c'est un critère d'appréciation convoqué par les juges qui cherchent à associer à un article une vidéo au contenu proche.

D'un autre côté, les enjeux économiques du projet font que la pertinence logique ne peut être l'unique facteur mobilisé pour apprécier une vidéo par rapport à un article. L'objectif étant de maximiser le nombre d'intégration vidéo, pour optimiser les revenus, explique qu'une vidéo peu voire non pertinente peut être retenue pour former une paire avec un article. Une telle vidéo est en effet utile d'un point de vue économique, puisqu'en étant présente en bas de la page de l'article, elle peut être lue par les internautes et augmenter le revenu de Mediabong. Il reste qu'une vidéo pertinente a évidemment plus de chances d'être lancée par les internautes qu'une vidéo sans lien apparent avec le contenu de l'article.

Dans notre travail, nous avons remarqué qu'une distinction était à faire entre les articles relatant des sujets d'actualité, et ce aux sujets plus spécialisés sur un thème particulier, hors de l'actualité, telle que la cuisine, la beauté, le jardinage, *etc.* Les articles

9. Nous rappelons ici que c'est le **texte associé à la vidéo** que nous analysons, et non son contenu audiovisuel.

d'actualité, largement majoritaires dans nos données, semblent plus facilement trouver des vidéos pertinentes que ceux traitant de sujets plus spécialisés. Ceci relève selon nous de deux facteurs. Le premier est que, par définition, un événement qui fait l'actualité est repris et développé par la majorité des diffuseurs de contenus et donc illustré par bon nombre de producteurs de vidéos. Ces articles-ci auront de fortes chances de trouver dans la collection plusieurs vidéos pertinentes, *i.e.* portant sur le même sujet bien défini.

La seconde raison tient selon nous au fait que les articles d'actualité décrivent régulièrement des événements d'actualité, ou en tout cas des sujets relatifs à un événement particulier. Bien que la notion d'événement soit largement débattue au sein de la communauté scientifique, notamment en linguistique, la définition de ce qu'est un *événement d'actualité* semble en revanche plus consensuelle. Dans la tâche de la détection automatique d'événement d'actualité (*New Event Detection*), issue du domaine plus large de *Topic detection and Tracking* (TDT), un événement est considéré comme "*quelque chose qui arrive sur un lieu particulier, à un moment donné*" (Allan, Papka et al., 1998). Cette définition simple, peut-être réductrice, permet toutefois de reconsidérer l'événement d'actualité comme un objet bien précis, dont il est possible de dégager une structure. En effet, les éléments de la définition qu'en donne le TDT permettent de dégager trois dimensions pour représenter l'événement d'actualité, répondant aux questions classiques exploitées dans l'écriture de presse – *Quoi ? Où ?* et *Quand ?* – auxquelles s'ajoute celle relative aux acteurs impliqués, c'est-à-dire *Qui ?* (Adam, 1997).

En se positionnant dans ce paradigme, on peut reconsidérer un événement comme un objet à quatre dimensions, dont les valeurs sur ces dimensions sont les réponses à chacune des précédentes questions. Même si un individu peut accorder plus d'importance à la dimension *Qui ?* qu'à la dimension *Où ?*, ou un autre plus d'importance à la dimension *Quand ?* qu'à toutes les autres, la représentation d'un événement d'actualité reste assez délimitée. Nous pensons que cette représentation particulière participe au fait que les juges soient globalement toujours d'accord sur l'appréciation des vidéos en réponse à des articles relatant des sujets d'actualité. C'est en tout cas ce que nous avons pu observer lors de nos expérimentations.

En revanche, pour les articles aux sujets plus spécialisés, aucune structure telle que celle-ci ne peut être dégagée. Même s'ils sont moins fréquents que ceux d'actualité, certains des partenaires diffuseurs sont spécialistes des contenus de beauté, santé, cuisine, décoration, sciences, ou encore nouvelles technologies. Si certaines vidéos peuvent être créées par les partenaires producteurs sur certains des sujets abordées par ces articles, tel n'est pas toujours le cas. De ce fait, comment garantir une intégration vidéo dans l'article, tout en sachant qu'aucune n'est vraiment pertinente dans la collection ?

En observant les jugements d'utilisateurs sur un échantillon d'articles, nous avons tenté d'éclaircir ces interrogations. Afin de comparer différentes configurations de système

d'appariement article-vidéo, nous avons mené une série d'expérimentations détaillées au Chapitre 6. Dans ce but, un jury de trois évaluateurs à été recruté pour évaluer manuellement des paires article-vidéo et nous disposons alors de données révélant leurs jugements de pertinence. Les exemples cités par la suite sont extraits de l'ensemble de test construit pour ces expérimentations, dont les titres des 50 articles le composant est présentée en Annexe D.

Nous distinguons dans cet ensemble deux catégories d'articles. La première regroupe les articles d'actualité, qui ne décrivent pas nécessairement des événements d'actualité, mais qui relatent au moins des sujets relatifs à un événement particulier. Tel que nous l'évoquions précédemment, ces articles sont largement sur-représentés [10] dans les données que nous traitons quotidiennement. L'ensemble de test contient 33 articles traitant d'actualité, soit 66%, et 17 articles spécialisés, soit 34%.

C'est dans un souci de lisibilité que nous ne présentons ici que les titres des articles pour illustrer nos propos, mais c'est bien sur l'ensemble de leurs contenus textuels (titre et description) que nous avons travaillé [11]. Toutefois, cette remarque est l'occasion de préciser que le titre d'un article de presse est généralement saillant par rapport au corps de l'article, grâce à différents types d'indice : d'une part, il occupe une position initiale et détachée par rapport au reste du texte, ce qui participe à sa mise en relief ; d'autre part, *« les titres de Une obéissent à [une] même logique : condenser l'information, attirer l'attention »* selon (Ringoot, 2014). En ce sens, ils présentent généralement de manière claire l'essentiel de l'information détaillée ensuite dans l'ensemble de l'article.

Concernant les articles d'actualité, on observe que les utilisateurs ont tendance à préférer les vidéos récentes, faisant intervenir l'acteur ou l'un des acteurs cités dans l'article [12]. Plus particulièrement, ce sont les noms de personne qui guident les utilisateurs dans l'appréciation des résultats. Régulièrement, on observe qu'une vidéo qui n'évoque pas nécessairement le sujet de l'article mais un autre impliquant le même acteur, est considérée comme bonne pour former une paire. Autrement dit, c'est régulièrement la dimension *Qui ?* qui est surpondérée dans les représentations d'articles à caractère événementiel. Dans une représentation vectorielle, la pondération des dimensions traduit l'importance accordée à celles-ci. Lorsque l'on dit qu'une dimension, en l'occurrence *Qui ?*, est surpondérée par rapport aux autres, cela signifie qu'on lui accorde une plus grande importance dans la représentation de l'article.

10. Bien que la sélection des requêtes de l'ensemble de test soit aléatoire, nous y avons ajouté une condition afin de conserver une répartition représentative des contenus selon leurs thématiques. C'est pour cette raison que l'on observe une majorité d'articles d'actualité qui sont effectivement sur-représentés en conditions réelles.

11. *cf.* Annexes A et B présentant des illustrations d'articles et de vidéos du corpus.

12. Ceci va dans le sens de l'affirmation de (Landragin, 2004) concernant le fort facteur de saillance que constitue le fait d'être un nom propre.

De nombreuses études sur le discours médiatique (Moirand, 2007 ; Steimberg, 2012) viennent étayer ces observations. Leurs auteurs ont démontré le lien entre dénominations et représentations communes en soulignant notamment l'importance des mentions de personnes et de lieux, ancrant l'actualité relatée dans un réel social. Le fait de nommer, ou plus précisément de figer la dénomination d'un référent particulier, participe à la construction d'un savoir partagé qu'un lecteur est capable d'identifier rapidement comme tel et qui se pose alors à lui comme élément saillant par rapport au reste du texte.

Ainsi, pour l'exemple **(1)** ci-dessous, toutes les vidéos au sujet de *Jacqueline Sauvage* ont été jugées pertinentes par l'ensemble des évaluateurs, et non simplement celles au sujet de la demande de grâce par ses filles. Cet article s'inscrit en fait au sein d'une sorte de dossier d'actualité au sujet de *l'affaire Jacqueline Sauvage*. Dans des cas comme celui-ci, nous avons observé que les utilisateurs ont toujours tendance à accepter les vidéos décrivant des faits relatifs au dossier, indépendamment du sujet particulier décrit dans l'article en cours d'analyse.

Le même constat est fait quant aux résultats de l'exemple **(2)**. *Fidel Castro* constitue l'élément principal du titre de l'article de par sa nature de nom propre, d'avantage souligné par son rejet en fin de proposition. En revanche ici, l'événement décrit est ponctuel et ne s'intègre pas dans une *affaire*. Les utilisateurs se sont donc ici accordés sur le jugement positif des vidéos au sujet des funérailles de *Fidel Castro* mais ont rejeté celles évoquant ce même personnage dans d'autres contextes.

(1)
Les filles de Jacqueline Sauvage ont déposé une demande de grâce totale à l'Elysée.

(2)
Discrétion et sobriété pour les funérailles de Fidel Castro

(3)
Bolloré visé par une enquête en Italie dans l'affaire Mediaset

(4)
Législatives : mécontent du candidat investi par Les Républicains, Patrick Balkany fait retirer les affiches de François Fillon à Levallois

Concernant l'exemple **(3)**, aucune des vidéos 242 proposées n'évoque *Vincent Bolloré* au sein de l'affaire *Mediaset*. Toutefois, 41 vidéos sur 43 au sujet de *Bolloré* ont été appréciées positivement. Par ailleurs, 5 des vidéos proposées relatent l'affaire *Mediaset* en citant *Vivendi* comme acteur principal. Parmi celles-ci, 4 ont été unanimement jugées pertinentes. *Vincent Bolloré* étant patron de *Vivendi*, un lien sémantique unit ces deux entités. Les vidéos impliquant cette société en tant qu'acteur, sujet de l'action décrite, sont considérées comme similaires à l'article. C'est donc ici plutôt le rôle sémantique qui

a été privilégié pour lier les vidéos à l'article.

En revanche, sur ce même exemple, le facteur de récence sur lequel MEDIABONG a particulièrement insisté en début de projet ne semble pas influencer les jugements. Certaines des vidéos remontant à plus de deux ans avant la publication de l'article ont été appréciées positivement. Cela dit, l'affaire *Mediaset* en question s'étend sur plusieurs années et il n'est donc pas aberrant d'associer une ancienne vidéo, relatant un fait relatif à ce dossier de rachat. La lecture de la vidéo par les internautes consultant la page de l'article peut alors constituer une sorte de rappel des faits précédents.

À l'inverse, dans l'exemple **(4)**, très peu de vidéos ont été jugées correctes par l'ensemble des trois juges. Sur 214 vidéos présentées, 32 relataient des faits impliquant Patrick Balkany. Or seulement deux d'entre elles ont été unanimement jugées correctes. La première, bien que relatant un sujet différent de celui évoqué dans l'article, positionne Patrick Balkany comme acteur et a été produite le même jour que celui de la publication de l'article. La seconde en revanche n'est pas liée à l'article quant au sujet décrit, et sa production remonte à plus d'un an avant la publication de celui-ci [13]. Or c'est également le cas pour une vingtaine d'autres vidéos proposées qui ont pourtant été rejetées par les juges. Notamment une vidéo très similaire à cette dernière, relatant la saisie de biens immobiliers de Patrick Balkany et datant d'août 2015, a été jugée non pertinente relativement à l'article.

Cet exemple démontre selon nous la difficulté de juger de la qualité d'une paire article-vidéo sans définition stricte de critères formels. Nous pourrions par exemple envisager un protocole présentant des règles du type "une vidéo doit partager au moins les valeurs des dimensions *Qui ?* et *Où ?* de l'article pour être considérée comme pertinente". Cela systématiserait le jugement et en faciliterait la modélisation. Toutefois, n'étant parvenu à ces considérations qu'en fin de thèse, après observation et analyse récurrentes des jugements de pertinence, aucun protocole de ce type n'a été mis en place pour l'évaluation des résultats du système d'appariement développé.

En l'absence de protocole d'évaluation, les juges sont donc libres d'apprécier les vidéos sur les critères qu'ils souhaitent et ils semblent ne pas toujours mobiliser les mêmes. La tâche d'attribution de jugement étant fastidieuse, ce comportement pourrait être mis sur le compte de la fatigue des utilisateurs qui saturent après un certain temps de travail (MA et al., 2016). Ce fait complexifie d'avantage la tâche qui est la nôtre puisque pour construire un système automatique, il est nécessaire de bien saisir en amont ce que ses utilisateurs souhaitent obtenir en sortie. Or il semble ici qu'eux-mêmes parviennent difficilement à délimiter ce besoin.

La série d'exemples qui suit présente des titres d'articles typiques dans la presse écrite,

13. Elle relate l'implication de Patric Balkany dans une affaire de fraude fiscale, et date d'octobre 2015.

utilisant les *deux-points* pour mettre en relief une certaine information. Sur les données de notre ensemble de test, aléatoirement sélectionnées, ces structures de titres ne représentent pas moins de 65% des cas. Une étude des années 90 sur les titres de presse intégrant cette ponctuation postule qu'une telle structure correspond à une configuration de type thème/rhème (Bosredon et al., 1992). Or la position de thème par rapport à celle de rhème figure parmi les critères sur lesquels repose la saillance linguistique (Landragin, 2011).

(5)
Brexit : May souhaite une sortie aussi "en douceur que possible"

(6)
César : Huppert, Cotillard, Ulliel, Sy parmi les nommés

(7)
Strasbourg : Proprios qui louent dans l'illégalité des appartements « type Airbnb »

(8)
Corse : Ils veulent rencontrer Christophe Maé, leur message cartonne sur les réseaux sociaux

Nos observations ne vont pourtant pas toutes dans cette direction-là. Si pour les exemples **(5)** et **(6)**, l'unité mise en exergue avant les deux points constitue bien le sujet principal de l'article, tel n'est pas le cas pour les deux exemples suivants, **(7)** et **(8)**. Pour les deux premiers en effet, l'ensemble des vidéos proposées respectivement au sujet du *Brexit* et de la *cérémonie des Césars* ont été jugées correctes. En revanche, en **(7)**, c'est l'entité *Airbnb* qui a été considérée comme saillante par les juges qui ont largement apprécié les vidéos impliquant cette entreprise. De la même façon en **(8)**, les juges se sont focalisés sur le chanteur *Christophe Maé* en notant positivement les vidéos à son sujet. Dans ces deux précédents cas, l'unité mise en exergue par les deux-points participe à la localisation du fait relaté mais n'en constitue pas l'information principale.

Ce constat tend de nouveau à souligner la force de la saillance sémantique portée par les noms propres, notamment les noms de personnes et d'entreprises. Par conséquent, lorsqu'aucun nom de personne ou d'entreprise n'est mentionné dans l'article, tout du moins dans son titre, quelles sont les unités que les juges considèrent comme saillantes ?

Lorsqu'il s'agit d'informations d'actualité internationale, la localisation est régulièrement mise en avant, comme en **(9)**, **(10)** et **(11)**. En plus de constituer des noms propres de pays, on observe dans ces exemples trois différents indices de saillance linguistique. En **(9)** et **(10)** l'indice est typographique, avec l'utilisation de *deux-points* pour le premier et de la *virgule* pour le second. Quant au dernier, en **(11)**, c'est la position sujet qui permet à l'entité *Nigeria* d'être saillante.

(9)
Famine au Soudan du Sud : le président promet d'aider l'accès aux ONG

(10)
À Alep, les civils doivent aussi faire face à la soif, à la faim et au froid

(11)
Le Nigeria ouvre une enquête après un bombardement qui a fait 70 morts

Dans ces cas-là, les juges apprécient globalement toutes les vidéos relatant des faits s'étant déroulés dans le pays évoqué dans l'article. Du point de vue de la représentation formelle des événements, c'est à la dimension *Où ?* que le plus fort poids est accordé.

Lorsqu'aucune information sur les lieux ou les acteurs impliqués n'est disponible, c'est alors sur les informations relatives à la dimension *Quoi ?* de l'événement que la saillance opère. En **(12)**, **(13)** et **(14)**, on observe des cas où une série d'événements a amené à créer une *affaire* dans le paysage médiatique. Or dans la presse, les références à un événement qui s'étend ont tendance à s'homogénéiser jusqu'à former une expression figée (Steimberg, 2006 ; Krieg-Planque, 2009 ; Battistelli et al., 2014). Ce procédé de nominalisation d'événement participe à son identification claire et univoque dans l'inconscient collectif. Ainsi, lorsqu'un titre de presse débute par la mention *moteurs diesel* ou *grippe aviaire*, qui plus est mise en avant typographiquement, cela participe à identifier le contexte particulier dans lequel s'inscrivent les faits.

(12)
Moteurs diesel **:** des juges d'instruction vont enquêter sur Renault

(13)
Violence dans les abattoirs **:** les députés refusent les caméras

(14)
Grippe aviaire **:** la justice ouvre une enquête après le scoop sur l'abattage des canards du Sud-Ouest

Il est donc particulièrement important de savoir repérer ces expressions nominales, et de les intégrer sous cette forme-ci dans la représentation des contenus. En segmentant brutalement les textes, pour ne conserver que les mots indépendamment les uns des autres, on perdrait les précieuses informations sémantiques que portent ces syntagmes.

Pour clore cette série d'illustrations, nous revenons sur les contenus atemporels qui ne s'inscrivent pas dans l'actualité médiatique. Le comportement des juges vis-à-vis de ces contenus particuliers est manifestement dépendant de la disponibilité de vidéos relatives aux sujets qui y sont décrits.

Dans les cas où l'information saillante est une entité bien définie, telle que *Facebook* en **(15)** et *Miss France* en **(16)**, les juges semblent accepter tous les contenus impliquant

ces entités-ci.

> **(15)**
> Si vous utilisez Facebook, vous vivrez potentiellement plus longtemps
>
> **(16)**
> Miss France : Les rêves d'Alicia ? Une famille «soudée», du mannequinat et «Touche pas à mon poste»
>
> **(17)**
> "Quartiers populaires" ? Mais que de mensonges derrière ces mots !
>
> **(18)**
> La recette du chia pudding d'Olivia

En **(17)**, l'expression *quartiers populaires* est clairement saillante, de part sa structure-même, sa position initiale et sa mise entre guillemets. Toutefois, le sujet de l'article est large et constitue plus une discussion au sujet des quartiers populaires que la description d'un fait particulier les concernant. Il s'agit ici d'un cas typique pour lequel aucune vidéo de la collection MEDIABONG n'est à même d'illustrer le contenu de l'article. En revanche, nous supposions que des vidéos impliquant des quartiers populaires auraient pu être appréciées positivement par les juges qui souhaitent associer une vidéo aux articles dans un maximum de cas. Or parmi l'ensemble proposé pour celui-ci, 5 vidéos mentionnent les quartiers populaires, mais seul un juge sur les trois impliqués les a considérées comme de bonnes candidates pour l'appariement à l'article. Les deux juges mécontents ont en fait rejeté l'ensemble des vidéos proposées pour cet article. Ils ont donc préféré ne rien lui associer plutôt que d'y intégrer une vidéo qu'ils estimaient trop éloignée thématiquement.

À l'inverse dans l'exemple **(18)**, bien qu'aucune vidéo de la collection ne décrive exactement la recette du *chia pudding*, il en existe plusieurs présentant des recettes de *pudding*. Toutes celles sur ce sujet présentées aux juges ont unanimement été considérées comme bonnes pour illustrer l'article.

Cette illustration de la tâche d'appariement qui nous a été confiée chez MEDIABONG souligne les difficultés qu'elle présente. L'une d'elles tient au fait que tous les articles traités comme requêtes par le système ne sont pas également exigeants quant à la sortie attendue. Lorsque le sujet qu'il relate est parfaitement défini, la vidéo appariée doit y être similaire sur un maximum de dimensions, si ce n'est toutes. En revanche, lorsque le sujet relaté est plus vaste, ou moins bien défini, la vidéo assortie peut se satisfaire d'une similarité sur certaines dimensions seulement. Cette différence d'attente relève simultanément de l'incomplétude de la collection de vidéos interrogée et du contexte économique nécessitant une intégration vidéo dans le plus de cas possibles. S'y ajoute la subjectivité du jugement de pertinence d'un appariement, puisque les critères de pertinence d'un appariement ne peuvent être clairement définis.

Face à cela, nous devions faire des compromis, en acceptant notamment de *négliger* la pertinence afin de favoriser le retour d'au moins un résultat. Il s'agissait d'un juste équilibre à trouver, car nous ne souhaitions pas non plus proposer des appariements aberrants.

Conclusion

Dans cette première partie, nous avons discuté de la notion d'*appariement*, au cœur de nos problématiques de recherche. Si intuitivement, le fait de former une paire ne semble pas insurmontable, nous avons témoigné malgré tout des difficultés que pouvaient présenter la tâche de jugement qualitatif d'une paire d'objets.

Nous sommes d'abord revenus sur le principe de *similarité*, qui sous-tend la tâche d'appariement. Nous avons montré que former une paire d'objets nécessite en amont de définir certains critères, certaines dimensions sur lesquelles les objets comparés doivent se rapprocher, voire s'équivaloir. Après avoir introduit quelques généralités sur la similarité, nous nous sommes attardés sur la comparaison d'objets particuliers que sont les textes. Ils constituent en effet notre objet d'étude et nous avons discuté de la complexité qui les caractérisent, essentiellement due à la diversité des dimensions qui les composent. De cette complexité découle celle relative à la façon de mesurer la similarité entre deux textes. Nous avons alors présenté différentes approches de la littérature abordant ces problématiques et des propositions pour leur résolution.

Nous sommes ensuite revenus en détails sur les critères impliqués dans l'appréciation d'un résultat fourni par un SRI. Nous nous sommes interrogés sur la façon dont les utilisateurs de tels systèmes évaluaient les résultats fournis en réponse aux requêtes qu'ils formulaient. En nous positionnant dans le contexte des sciences de l'information, nous avons montré que la pertinence était régulièrement proposée en réponse à ces interrogations. Toutefois, il est apparu que la définition de cette notion était plus que difficile, puisqu'elle caractérise différentes propriétés des résultats, pas nécessairement compatibles.

La discussion s'est achevé par la mise en perspective des précédentes notions avec les données de notre corpus. À partir d'exemples d'appariements article-vidéo fournis automatiquement, nous avons observé les jugements délivrés par un jury d'évaluateurs à leur égard. Nous souhaitions observer les caractéristiques des textes sur lesquels se basaient ces utilisateurs pour juger de la pertinence d'une vidéo relativement à un article. Nous souhaitions également savoir si certaines d'entre elles, considérées comme saillantes d'un point de vue linguistique, constituaient des dimensions plus importantes que les autres dans les représentations que les utilisateurs se font des textes. En d'autres termes, nous désirions comprendre comment l'appariement article-vidéo était envisagé du point de vue

des utilisateurs afin d'optimiser en ce sens l'algorithme implémenté pour résoudre cette tâche.

Toutefois, les conclusions de ces observations ont contribué à soulever davantage de questions qu'elles n'en ont résolues. Nous avons malgré tout pu en tirer un enseignement relatif au système commandé par MEDIABONG, celui de trouver un compromis entre les différentes exigences caractérisant leur besoin initial et c'est ce que nous sommes efforcés de faire dans le développement du système proposé dans cette thèse.

Deuxième partie

État de l'art

Chapitre 1

Recherche d'information

La recherche d'information est une tâche née du besoin de retrouver efficacement des données non structurées dans une collection documentaire, en réponse à un besoin d'information spécifique. C'est Calvin Mooers qui, en 1948, propose pour la première fois dans son mémoire de Master le terme d'*information retrieval*, avec la définition suivante, reprise en 1989 dans la seconde édition de l'Oxford English Dicrtionnary : *"The requirement of information retrieval of finding information whose location and very existence is a priori unknown... "*[1] (Mooers, 1948). Lorsqu'il a fallu, en une expression claire et concise, traduire ce concept en français, le choix de *recherche d'information* s'est imposé. Il reste néanmoins discutable du fait que cette tâche particulière n'est pas qu'une simple recherche, c'est un processus complet partant d'un besoin qui déclenche en effet une recherche aboutissant elle-même à un résultat, qu'il satisfasse ou non le besoin initial.

C'est ensuite dans les années 1950, après la seconde guerre mondiale qui fut riche en création documentaire, ainsi qu'avec le développement de collections de documents spécialisés, que la nécessité d'automatiser ce processus a réellement émergé. L'idée d'utiliser les ordinateurs, dont les capacités de calculs dépassent celles des hommes, s'est rapidement imposée. Les premiers travaux d'expérimentations en la matière ont débuté dès la fin des années 1950, l'un des plus connus et encore cités aujourd'hui étant le projet Cranfield (Cleverdon, 1967).

Dans ce chapitre, nous commençons en Section 1.1 par présenter plus en détails la tâche de recherche d'information (RI) et ses objectifs. Nous développons dans les sections suivantes les différentes étapes de l'implémentation d'un tel processus, en débutant par la tâche d'indexation des contenus (Section 4.2), puis celle du calcul de similarité entre requête et documents (Section 1.3). Nous exposons finalement en Section 1.5 les méthodes classiques d'évaluation des systèmes de recherche d'information (SRI), en présentant succinctement les campagnes d'évaluation réputées dans le domaine.

1. *La nécessité de la recherche d'information [est] de trouver des informations dont la localisation et l'existence-même sont a priori inconnues.*

1.1 Présentation de la tâche

Avant d'être une discipline de l'informatique, la RI est d'abord une discipline des sciences de l'information. Son objet d'étude relève en effet de la satisfaction d'un utilisateur humain quant à un besoin d'information formulé et il est donc nécessaire d'analyser en amont de tout développement technique les problématiques que cet objet impose. Par ailleurs, un autre défi de la RI, par rapport à toute autre discipline de l'informatique, tient à la complexité des données qui lui sont données à traiter, à savoir des données textuelles, non structurées.

Partant de ces contraintes initiales, les problématiques de la RI relèvent de différents aspects. Dans le cadre de cette thèse, on s'interroge notamment sur celles concernant la représentation des contenus textuels de sorte qu'un système soit capable d'en *comprendre* le sens. Quels traitements est-il nécessaire d'imposer aux documents textuels pour en créer une modélisation avec laquelle une machine serait en mesure d'interagir ? Comment saisir automatiquement la complexité dont les textes sont empreints, alors même que la tâche peut être délicate pour un humain ? Découlent de ces interrogations sur les documents une autre problématique relevant plus des sciences humaines : comment comprendre, puis représenter, le besoin d'information de l'utilisateur qui fournit au système une requête censée rendre compte de ce besoin ? On peut se demander s'il n'est pas réducteur de traduire un besoin d'information qui prend forme dans l'esprit d'un individu en un ensemble de mots-clés que l'on soumet à un moteur de recherche. La question sous-jacente serait alors de savoir comment traduire un besoin d'information, qui peut s'avérer complexe, en un langage simple qu'un ordinateur puisse appréhender.

Discipline issue de la recherche documentaire, la RI se donne pour objectif de trouver des documents, régulièrement textuels donc non structurés, satisfaisant un besoin d'information donné, parmi un ensemble de documents composant une collection. En recherche documentaire, cette activité était réservée à des spécialistes de la documentation qui maîtrisaient les langages d'interrogation des systèmes développés. Mais le monde a aujourd'hui évolué et l'accès à l'information s'est particulièrement démocratisé avec l'avènement du Web 2.0 dans des années 2000, grâce auquel l'utilisateur est devenu un acteur central des outils de recherche mais aussi de création et de diffusion de l'information. Ce changement de paradigme a entraîné de sérieuses modifications d'approches, à tous les stades du processus de RI que nous développons dans ce chapitre, du fait d'un considérable changement d'échelle : s'il était envisageable, pour une collection documentaire de quelques milliers d'entrées, d'indexer manuellement les documents, l'entreprise devient inconcevable lorsque l'on considère une collection de plusieurs centaines de milliers de documents, voire de quelques millions. Par ailleurs, les requêtes étaient précédemment formulées par des experts de systèmes documentaires, dans un langage et avec un voca-

bulaire contrôlé. Aujourd'hui, elles sont soumises par des utilisateurs *lambda*, loin d'être spécialistes des domaines sur lesquels ils s'interrogent, qu'ils peuvent qui plus est formuler en langue naturelle, non contrôlée.

Loin d'être exhaustif, au regard de l'abondante littérature dans le domaine depuis sa naissance et plus que prolifique aujourd'hui encore, ce chapitre s'attache à la présentation d'une synthèse des méthodes proposées dans l'état de l'art de la RI. Nous insistons notamment sur les sous-tâches du processus global auxquelles nous nous sommes particulièrement intéressés dans le cadre de cette thèse, à savoir la représentation des contenus, les modèles de similarité entre contenus et l'évaluation des systèmes.

1.2 Indexation des contenus

L'étape préliminaire en RI consiste à représenter les contenus textuels de façon à ce qu'ils soient intelligibles par une machine qui n'est pas en mesure d'interpréter un texte brut. L'idée est de réduire un document à un ensemble de mots-clés représentant le sens global que son contenu véhicule. Ces mots-clés peuvent correspondre à des termes simples (*i.e.* monolexicaux) ou composés (*i.e.* polylexicaux), voire des phrases dans certains cas, extraits du texte-même ou qui lui sont assignés sans y figurer textuellement.

Afin d'homogénéiser l'indexation de toutes les entrées d'une base documentaire et permettre leur comparaison, il est nécessaire de convenir d'un vocabulaire d'indexation (ou langage d'indexation) qui peut être libre ou contrôlé. L'indexation contrôlée est privilégiée dans le cas de collection documentaire spécialisée, pour laquelle on fournit à l'indexeur une liste figée de termes d'indexation grâce auxquels tous les documents de la collection peuvent être indexés. À l'inverse, dans le cas de collections plus hétérogènes, on privilégie une indexation libre, sans liste fermée, permettant d'assigner à un document tous les termes souhaités.

Si les notions de *descripteurs* et de *termes d'indexation* sont globalement synonymes, la première est régulièrement utilisée dans le cadre d'une indexation contrôlée et/ou dans celui d'un système de recherche documentaire, tandis que la seconde est plutôt employée dans le cadre d'une indexation libre et/ou dans celui d'une RI à plus grande échelle. Nous emploierons donc par la suite la notion de *termes d'indexation* qui rend compte de ce que nous présentons dans cette section, à savoir les processus en jeu dans une indexation libre et automatisée.

1.2.1 Sélection des termes

Cette sous-section présente un éventail de possibilités d'attribution de termes d'indexation aux documents d'une collection. La Table 1.1 présente un exemple de petite collection documentaire, que nous noterons désormais *collection-exemple*, grâce à laquelle nous illustrerons chacune des méthodes d'indexation développée ici. Le vocabulaire d'indexation, lorsqu'il est libre, correspond à l'ensemble des termes distincts résultant de la phase d'indexation d'une collection documentaire.

Doc 1 :	Financement libyen, révélations de Buisson... la mauvaise passe de Sarkozy
Doc 2 :	Libye : 4 morts dans l'attaque d'un canot de migrant par des hommes armés
Doc 3 :	Migrants à Calais : La Belgique sur ses gardes avant le démantèlement du camp
Doc 4 :	Un nouveau document libyen mentionne le financement de la campagne Sarkozy en 2007
Doc 5 :	L'affaire Bygmalion, de Copé à la campagne Sarkozy
Doc 6 :	Éleveurs : Stéphane Le Foll se rendra finalement à Caen cet après-midi

TABLE 1.1 – Un exemple de collection documentaire

Index brut

L'indexation la plus élémentaire que l'on puisse envisager est celle consistant à retenir l'ensemble des termes simples composant le texte du document à indexer, en considérant l'espace ou tout autre caractère non alphanumérique comme séparateur de termes. Pour homogénéiser un tant soit peu le vocabulaire généré, il est d'usage de normaliser les termes en casse. Ainsi, si un terme apparaît dans un document en majuscule et dans un autre document en minuscule, il sera ramené dans le vocabulaire à sa forme minuscule.

La Table 1.2 illustre le résultat d'une telle indexation appliquée aux documents de la *collection-exemple*.

Doc 1 :	[buisson, de, financement, la, libyen, mauvaise, passe, révélations, sarkozy]
Doc 2 :	[4, armés, attaque, canot, d', dans, de, des, hommes, l', libye, migrant, morts, par, un]
Doc 3 :	[à, avant, belgique, calais, camp, démantèlement, du, gardes, la, le, migrants, ses, sur]
Doc 4 :	[2007, campagne, de, document, en, financement, la, le, libyen, mentionne, nouveau, sarkozy, un]
Doc 5 :	[à, affaire, bygmalion, campagne, copé, de, l', la, sarkozy]
Doc 6 :	[à, après, caen, cet, éleveurs, finalement, foll, le, midi, rendra, se, stéphane]

TABLE 1.2 – Exemple d'indexation *brute* des contenus

Nous observons tout d'abord que l'ordre linéaire du texte est perdu, les termes sont en effet considérés comme des constituants indépendants du texte qu'ils composent en fait ensemble. C'est une représentation textuelle qui est dite en *sac de mots*, expression dont

l'image reflète bien l'idée d'un ensemble non ordonné. Nous reviendrons par la suite sur cette contrainte forte, susceptible d'entraîner des erreurs de représentation documentaire dont peuvent découler ensuite des erreurs dans le processus de recherche d'information et dans ses résultats.

Une seconde remarque concerne la fréquence d'occurrence des termes. Prenons l'exemple du **Doc 1**, où l'on trouve deux fois la préposition *de* dans le texte brut, tandis qu'elle ne figure qu'une seule fois dans l'index extrait. Ce que l'on conserve à l'issue de l'indexation sont en fait les types – *i.e* les notions génériques – et non les tokens – *i.e.* les occurrences particulières de types (Peirce, 1931). Dans le **Doc 1**, nous avons donc pour la préposition *de* deux tokens mais un unique type. Néanmoins, la fréquence d'occurrence d'un terme dans un document est un facteur clé dans l'indexation, intervenant lors de la phase d'attribution de poids aux différents termes sur laquelle nous reviendrons en 1.3.2.

Nous constatons par ailleurs que la règle de segmentation rudimentaire appliquée, consistant à considérer tout caractère non alphanumérique comme séparateur, pourrait être affinée. L'index du **Doc 6** en illustre bien les limites, puisque le terme initial *après-midi* s'est vu transformer en deux termes d'indexation indépendants, *après* et *midi*. De la même façon, *Stéphane Le Foll* est un nom propre que l'on peut considérer comme une unité polylexicale particulière, dénotant une entité extra-linguistique de type personne. Cette unité voit ses composants séparés à l'issue de l'indexation et la référence à l'entité n'est alors plus possible dans une telle représentation.

Le vocabulaire résultant de l'indexation brute de cette collection documentaire, présenté en Table 1.3, compte 44 termes. Parmi ceux-ci, il apparaît clairement que tous ne sont pas des termes participant à la construction du sens des textes les intégrant. C'est notamment le cas des mots grammaticaux qui participent plus à la forme du texte qu'à son fond, *i.e.* l'information substantielle qui y est décrite. Par ailleurs, nous aimerions pouvoir faire comprendre à un système automatique que les termes *migrant* et *migrants*, respectivement présents dans les **Doc 2** et **Doc 3**, renvoient au même concept et ne varient que d'une flexion morphologique (en l'occurrence, le nombre).

Ces remarques démontrent les nombreuses faiblesses d'une indexation brute, dénuée de toute normalisation et de tout filtre. Sont donc présentées par la suite des méthodes d'indexation plus *intelligentes*, considérant les contraintes précédemment évoquées.

Prétraitements linguistiques

La première opération d'indexation est la segmentation du texte en unités, régulièrement appelée *tokenisation*. Se contenter d'une règle générique, telle que distinguer ces unités sur la base de simples séparateurs graphiques, est loin d'être satisfaisant, notamment parce qu'elle ne considère pas les spécificités de la langue analysée. (Palmer, 2000)

4	caen	des	la	par
2007	calais	document	le	passe
à	camp	du	libye	rendra
affaire	campagne	éleveurs	libyen	révélations
après	canot	en	mauvaise	sarkozy
armés	cet	finalement	mentionne	se
attaque	copé	financement	midi	ses
avant	d'	foll	migrant	stéphane
belgique	dans	gardes	migrants	sur
buisson	de	hommes	morts	un
bygmalion	démantèlement	l'	nouveau	

TABLE 1.3 – Exemple de vocabulaire issu de l'indexation *brute*

insiste sur le fait qu'un système de tokenisation doit s'appuyer sur la structure de la langue qui lui est donnée à segmenter. Chaque langue a des caractéristiques morphologiques et syntaxiques qui lui sont propres, en fonction desquelles les règles de segmentation varient.

Envisager de construire un outil de tokenisation automatique nécessite donc de considérer en amont les caractéristiques de la langue étudiée, qui peuvent former un système assez complexe de règles et d'exceptions. En français par exemple, nous souhaitons distinguer le déterminant du nom dans le syntagme *l'affaire*, pour obtenir deux tokens *l'* et *affaire*. Nous pourrions donc écrire une règle de segmentation spécifiant que deux chaînes de caractères séparées par une apostrophe doivent être disjointes. Seulement, en établissant une telle règle, nous obtiendrions des erreurs pour des composés lexicaux tels que *presqu'île* ou *aujourd'hui*. De la même façon, une règle pourrait signifier que deux chaînes de caractères alphabétiques séparées par un tiret, à l'instar d'*après-midi* ou de *soixante-douze*, devraient être considérées comme une seule et même unité. Mais une telle règle générerait des unités du type *pensez-vous* ou *a-t-il*, alors que les différents composants devraient ici être séparés.

Il existe pour le français de nombreux outils pour opérer un tel traitement. La bibliothèque NLTK [2] de Python propose un module de tokenisation avec de nombreuses options de segmentation et de langue. TREETAGGER (SCHMID, 1994), outil d'annotation linguistique, dispose également d'un module de prétraitement chargé de la segmentation du texte soumis en entrée.

Une fois le texte correctement segmenté, il convient d'en normaliser les unités afin d'obtenir pour chacun des documents d'une collection des représentations comparables. L'hétérogénéité du vocabulaire d'indexation est susceptible d'entraîner des erreurs dans le processus de RI. C'est le cas notamment des variantes orthographiques d'un terme, dénotant un même concept, qui seront distinguées dans le vocabulaire d'indexation si elles ne sont pas unifiées (***e.g.*** *événement* | *évènement* ; *alaise* | *alèze* | *alèse*).

2. http://www.nltk.org/

Afin de pallier ce biais, des processus de normalisation peuvent être mis en œuvre, dans le but de ramener sous une même forme canonique les variantes d'un même terme. (Amini et al., 2013) distinguent deux types de normalisations : les processus de normalisation textuelle et ceux de normalisation linguistique.

La normalisation textuelle est une transformation de surface qui unifie les graphies des termes pour en donner une représentation générique. Elle s'affranchit des alternatives formelles, telle que la casse ou l'accentuation, pour conserver les termes sous leur forme la plus élémentaire. Ces transformations sont toutefois à considérer avec précaution, car de la même façon que pour la segmentation brute sur les séparateurs graphiques, elles peuvent parfois intégrer plus d'ambiguïtés qu'elles n'en lèvent. Si l'on trouve par exemple dans un document la mention de l'entreprise *Orange* et que l'on normalise la casse du texte (tout en minuscule), une ambiguïté sera introduite du fait qu'*orange*, terme polysémique, deviendra terme d'indexation. Concernant la suppression des accents, mettons qu'un document contienne le terme *salé*, normalisé en supprimant l'accent : *sale* sera alors terme d'indexation et le sens du terme initial sera perdu.

La normalisation linguistique est une transformation plus profonde, unifiant les variantes morphologiques des termes. Il en existe deux niveaux. Le premier réduit à leur forme canonique non fléchie les variantes des termes d'une même catégorie syntaxique, il s'agit de la *lemmatisation*. Le second niveau réduit à une racine l'ensemble des termes, toutes catégories syntaxiques confondues, formés sur un radical commun : il s'agit de la *racinisation*, ou *stemming*. Plus précisément, l'opération de lemmatisation réduit un verbe à son infinitif (***e.g.*** *ai, eu, ayant = avoir*), un adjectif à sa forme masculin et singulier (***e.g.*** *beaux, belles = beau* ; *verts, vertes = vert*), et un nom à sa forme au singulier (***e.g.*** *sacs = sac* ; *animaux = animal*). Cela permet de réduire la taille du vocabulaire et fait apparaître des similarités entre des documents de la collection qui n'étaient initialement pas associés. C'est par exemple le cas des **Doc 2** et **Doc 3** de l'exemple initial, qui étaient complètement disjoints à l'issue de l'indexation brute mais qui partagent un terme du vocabulaire, *migrant*, après lemmatisation de leurs contenus (*cf.* Table 1.5). Pour parvenir à un tel résultat, les outils de lemmatisation – ou lemmatiseurs – procèdent à une analyse linguistique fine, contrairement aux outils de racinisation qui reposent globalement sur des règles de désaffixation dépendantes de la langue. L'idée générale en *stemming* est de retrouver la racine morphologique d'un terme, en conservant son radical porteur du sens et en éliminant les affixes. Par exemple en français, l'ensemble de termes [*migrez, migration, migrant, migrer, immigration, émigrantes*] ont tous la même racine *migr*. Si cette méthode permet une réduction encore plus drastique du vocabulaire d'indexation, elle est plus risquée que la lemmatisation de par sa capacité à sur-raciniser : les auteurs de (Moreau et al., 2006) donnent l'exemple de la racine *nat* qui représente à la fois les termes *nature* et *nation* qui n'ont pourtant aucun lien sémantique.

De ce fait, la racinisation est rarement privilégiée dans le traitement de contenus en français, pour lesquels la lemmatisation reste plus fiable. C'est en revanche une méthode qui continue d'être largement utilisée pour l'analyse de l'anglais, langue peu flexionnelle pour laquelle la lemmatisation est moins intéressante. C'est d'ailleurs sur cette langue que (Porter, 1980) a développé l'un des algorithmes de racinisation les plus populaires à base de règles et de listes de suffixes récurrents de la langue. Il existe malgré tout quelques outils de *stemming* pour le français, basés sur cet algorithme pionnier, tels que Carry (Paternostre et al., 2002) ou Snowball, outil libre accessible en ligne[3] dont la Table 1.4 propose une illustration des résultats.

Doc 1 :	[buisson, de, financ, la, libyen, mauvais, pass, rével, sarkozy]
Doc 2 :	[4, armé, attaqu, canot, d'un, dan, de, homm, l', liby, migr, mort, par]
Doc 3 :	[à, avant, belgiqu, cal, camp, démantel, du, gard, la, le, migr, se, sur]
Doc 4 :	[2007, campagn, de, docu, en, financ, la, le, libyen, mention, nouveau, sarkozy, un]
Doc 5 :	[à, affair, bygmalion, campagn, cop, de, l', la, sarkozy]
Doc 6 :	[à, apres-mid, caen, cet, éleveur, final, foll, le, rendr, se, stéphan]

Table 1.4 – Exemple d'indexation de contenus racinisés par Snowball

Concernant la lemmatisation, TreeTagger (Schmid, 1994) est indéniablement l'outil le plus connu et le plus utilisé par la communauté TAL, dont un exemple de sortie est présenté en Table 1.5

Doc 1 :	[buisson, de, financement, le, libyen, mauvais, passe, révélation, sarkozy]
Doc 2 :	[4, armé, attaque, canot, dans, de, du, homme, le, libye, migrant, mort, par, un]
Doc 3 :	[à, avant, belgique, calais, camp, du, démantèlement, garde, le, migrant, son, sur]
Doc 4 :	[2007, campagne, de, document, en, financement, le, libyen, mentionner, nouveau, sarkozy, un]
Doc 5 :	[à, affaire, bygmalion, campagne, copé, de, le, sarkozy]
Doc 6 :	[à, éleveur, après-midi, caen, ce, finalement, foll, le, rendre, se, stéphane]

Table 1.5 – Exemple d'indexation de contenus lemmatisés par TreeTagger

La dernière phase de la chaîne de traitement classique pour l'indexation des contenus en RI est le filtrage des termes. Nous remarquions en effet plus haut que le fait de conserver tous les termes des textes analysés était loin d'être pertinent, tous ne jouant pas un rôle lexical et/ou sémantique. Certains, dénotés à juste titre par l'expression de *mots-vides*, fonctionnent comme des outils de cohésion textuelle sans apporter d'information en tant que telle. Ces mots-vides faisant régulièrement partie de classes fermées, comme les déterminants ou les prépositions, il est d'usage d'en constituer une ressource à laquelle les indexes des documents sont comparés pour procéder à leur exclusion. Il s'agit d'un

3. http://snowball.tartarus.org/algorithms/french/stemmer.html

antidictionnaire, qui peut ne pas se limiter aux mots grammaticaux mais intégrer plus largement tous les termes d'une collection particulièrement présents dans les documents la composant. L'idée ici est qu'un terme apparaissant avec une forte fréquence documentaire, *i.e.* dans beaucoup de documents différents, est certainement moins discriminant pour représenter un document qu'un terme qui est plus rare. L'exemple le plus flagrant est celui des auxiliaires *être* et *avoir* qui, même s'ils sont considérés comme des mots lexicaux, n'aident pas à distinguer un document d'un autre puisqu'ils apparaissent dans presque tous les documents d'une collection. Par ailleurs, lorsque l'on travaille sur une collection documentaire spécialisée, les documents sont empreints de la terminologie du domaine étudié et certains mots lexicaux peuvent alors s'avérer peu discriminant : le terme *loi* dans un corpus juridique pourrait par exemple être écarté, en supposant qu'une grande majorité des textes le contiennent.

L'idée générale est donc de filtrer les mots grammaticaux pour ne conserver que les mots lexicaux, c'est-à-dire tous les noms, verbes, adjectifs et adverbes (Amini et al., 2013). Les termes de toutes les autres catégories devraient donc figurer dans l'antidictionnaire. Une solution alternative à ce type de ressource est de faire appel à des outils d'étiquetage en parties du discours (dits PoS pour *Part-of-Speech*) : si l'on connaît en effet les catégories des termes que l'on souhaite conserver, il suffit d'exclure tous ceux n'y correspondant pas. La procédure semble moins lourde que le filtrage par antidictionnaire qui implique dans un premier temps le recensement exhaustif des termes à exclure, puis dans un second temps la comparaison de chaque document à indexer à cette ressource. En revanche, comme toute opération automatisée, une telle annotation implique des risques d'étiquetage susceptibles d'exclure des termes qui appartiennent en fait à l'une des catégories à conserver ou, à l'inverse, de conserver des termes qui auraient dû être filtrés. Toutefois TreeTagger, également outil d'étiquetage morpho-syntaxique basé sur un apprentissage par arbre de décision probabiliste, atteint une précision de 96.14% sur le corpus de test du French TreeBank (Abeillé et al., 2003 ; Denis et al., 2010), révélant un faible taux d'erreurs.

Doc 1 :	[buisson, financement, libyen, mauvais, passe, révélation, sarkozy]
Doc 2 :	[armé, attaque, canot, homme, libye, migrant, mort]
Doc 3 :	[belgique, calais, camp, démantèlement, garde, migrant]
Doc 4 :	[campagne, document, financement, libyen, mentionner, nouveau, sarkozy]
Doc 5 :	[affaire, bygmalion, campagne, copé, sarkozy]
Doc 6 :	[éleveur, après-midi, caen, finalement, foll, rendre, stéphane]

Table 1.6 – Exemple d'indexation des contenus étiquetés par TreeTagger

La Table 1.6 présente le résultat d'une indexation des contenus de la *collection-exemple* tokenisés, lemmatisés, annotés en PoS par TreeTagger, puis filtrés pour ne conserver que les NOM, NAM, VER et ADJ [4]. Le vocabulaire d'indexation résultant de cette pro-

4. Soit, dans l'ordre, les noms communs, noms propres, verbes et adjectifs.

cédure, qui ne compte plus que 34 termes, est quant à lui présenté en Table 1.7.

affaire	camp	foll	mort
après-midi	campagne	garde	nouveau
armé	canot	homme	passe
attaque	copé	libye	rendre
belgique	document	libyen	révélation
buisson	démantèlement	mauvais	sarkozy
bygmalion	éleveur	mentionner	stéphane
caen	finalement	migrant	
calais	financement	migrer	

TABLE 1.7 – Exemple de vocabulaire d'indexation normalisé et filtré via TREETAGGER

Enrichissement sémantique

Les diverses méthodes de normalisation présentées permettent une sélection plus pertinente des termes d'indexation, et donc une meilleure représentation des contenus. Mais toutes, mises à part la tokenisation qui peut isoler certaines unités multi-mots, se contentent de considérer des termes simples, monolexicaux, en perdant alors toute la syntaxe et la sémantique induites par la linéarité des textes. Or ces expressions multi-mots (EMM) sont des unités lexicales qu'il faut considérer à part entière, puisqu'elles représentent souvent plus pertinemment les documents dans lesquelles elles figurent que les termes simples qui les composent. Étant le plus souvent non-compositionnelles (CONSTANT, 2012), il est nécessaire de les considérer comme un tout pour conserver leur sens initial qui n'est pas déductible de la somme de celui de ses composants. C'est notamment le cas des expressions formées d'unités lexicales contiguës, *i.e.* une suite immuable de mots comme les expressions figées, dont la tête peut être nominale (***e.g.*** *cordon bleu, tour de contrôle*), verbale (***e.g.*** *porter le chapeau*) ou adverbiales (***e.g.*** *tout à fait*), ainsi que les entités nommées (EN). Ces dernières forment des unités de sens faisant référence, dans leur définition initiale (EHRMANN, 2008), à des personnes (***e.g.*** *Amélie Nothomb, Stéphane Le Foll*), des lieux (***e.g.*** *New York, Tour Eiffel*) ou des organisations (***e.g.*** *Union Européenne, Université Paris Nanterre*). Notons qu'une EN n'est pas nécessairement multi-mots, elle peut n'être constituée que d'un terme simple, à l'instar de *Paris* ou *Berlin*, ou se réduire à un sigle comme c'est le cas d'un certain nombre d'organisations telles que l'*ONU* ou la *SNCF*. Toutefois, lorsqu'elles ne constituent qu'une unité simple, les EN sont correctement segmentées lors de la phase d'indexation et sont correctement extraites des documents dans lesquelles elles apparaissent. Ce qui reste en revanche un défi est le fait de les catégoriser en tant que telle.

La tâche d'extraction d'EN, particulièrement étudiée depuis que les conférences MUC [5]

5. *Message Understanding Conference*

s'y sont intéressées à la fin des années 80, se décompose en effet en deux grandes phases de traitement : la première consiste à les repérer dans un texte non structuré, la seconde à les typer. Deux approches, qui peuvent être combinées, sont généralement distinguées pour cette tâche : les approches linguistiques, et les approches probabilistes. Les premières se basent sur des analyses de surface, en établissant des règles d'extraction basée sur des marqueurs lexicaux régulièrement associés à des EN (***e.g. Monsieur*** *X ;* ***Le ministre*** *Y ;* ***Le lac*** *de Garde*) ou encore sur des patrons exploitant des informations morpho-syntaxiques et interrogeant des ressources externes telles que des dictionnaires de noms propres ou des ontologies. Les secondes approches sont basées sur l'apprentissage supervisé à partir de corpus annoté en EN. L'idée de ces méthodes est d'apprendre des règles d'extraction automatiquement, à partir des récurrences observées en corpus. Par exemple, si le système observe lors de l'apprentissage qu'après chaque occurrence de *Madame*, le segment qui suit est toujours annoté comme une EN de type PERSONNE, il en déduira une règle permettant de considérer *Madame* comme marqueur introduisant une EN. Grâce à cette règle, il sera en mesure de repérer les EN ainsi formées dans un nouveau corpus, non annoté.

Dans notre contexte de RI, nous nous intéressons au repérage des EN dans l'idée que nous ne souhaitons pas lier un document au sujet de *François Hollande* à un document au sujet de *François Fillon*, sur le seul fait qu'ils aient en commun le terme *François*. Toutefois, si nous proposons dans cette thèse un algorithme pour extraire ce type de segments (*cf.* Chapitre 4), l'étape de typage des candidats extraits se fait elle manuellement, part une équipe de salariés de MEDIABONG. La tâche globale de reconnaissance d'EN est donc, dans la version actuelle de notre système, semi-automatique.

1.2.2 Fichier inverse

L'indexation documentaire est en fait une phase de préparation à la RI, elle transforme les documents textuels en une représentation structurée. Lors de la phase de recherche, le système automatique interroge une base de données dans laquelle les indexes documentaires sont stockés. Parallèlement, une autre structure de données essentielle est alimentée, elle permet d'associer à chaque terme du vocabulaire d'indexation les documents de la collection le contenant : il s'agit du *fichier inverse*, ou *index inversé*. Le besoin d'information d'un utilisateur de SRI est régulièrement formalisé sous forme d'un ensemble de mots-clés constituant sa requête [6] et le fichier inverse permet de retrouver efficacement les documents de la collection contenant ces termes particuliers. À l'image des indexes que l'on peut trouver en début ou en fin de certains livres, cette structure permet d'accéder rapidement à l'information recherchée, en évitant dans ce cas au lecteur le parcours ex-

6. Nous reviendrons sur la notion de requête dans la section suivante (1.3) décrivant les modèles de RI.

haustif de toutes les pages de l'ouvrage pour ne se concentrer que sur celles contenant les concepts qui l'intéressent. Transposé à un SRI, ce fichier inverse permet de ne récupérer que les documents de la collection partageant au moins un terme avec la requête soumise. La forme minimale que cette structure peut adopter est une simple correspondance entre un terme et les documents qui lui sont associés (*cf.* Table 1.8). Elle peut néanmoins être plus informative et présenter également la fréquence documentaire du terme ou encore la position du terme dans chacun des documents le contenant (Amini et al., 2013).

Termes	**Documents associés**
affaire	**Doc 5**
après-midi	**Doc 6**
armé	**Doc 2**
attaque	**Doc 2**
belgique	**Doc 3**
buisson	**Doc 1**
bygmalion	**Doc 5**
caen	**Doc 6**
calais	**Doc 3**
camp	**Doc 3**
campagne	**Doc 4**
canot	**Doc 2**
copé	**Doc 5**
document	**Doc 4**
démantèlement	**Doc 3**
éleveur	**Doc 6**
finalement	**Doc 6**
financement	**Doc 1**, **Doc 4**
...	...

Table 1.8 – Illustration d'un fichier inverse, extrait de la *collection-exemple*

1.2.3 Pondération des termes

Nous évoquions plus haut l'importance de distinguer les termes en fonction de l'importance qu'ils ont dans la représentation d'un document. Cette importance peut être traduite par une pondération attribuée à chacun des différents termes de l'index documentaire. Cette section en présente les implémentations les plus classiques en RI, de la plus simpliste aux plus avancées.

Binaire

La méthode de pondération la plus basique est la pondération binaire. Il s'agit d'accorder un poids de 1 si un terme du vocabulaire d'indexation apparaît dans un document

donné et un poids de 0 pour tout terme du vocabulaire n'y figurant pas. Cette méthode simple permet d'isoler les termes du vocabulaire impliqué dans un document mais ne distingue pas ces termes entre eux puisque tous ont le même poids.

Fréquentielle

Dans ses travaux, (Luhn, 1958) suggère que la fréquence d'un terme dans un document reflète son degré de représentativité quant à l'information qui y est développée. Si l'auteur d'un document fait le choix d'utiliser plusieurs fois un même terme, c'est qu'il considère celui-ci comme particulièrement représentatif de l'information qu'il souhaite transmettre. En se basant sur cette hypothèse, considérer la fréquence brute, *i.e.* le nombre d'occurrences d'un terme dans un document (notée *TF* pour *Term Frequency*) comme son poids semble être pertinent.

TF*IDF

(Sparck Jones, 1972) considère qu'un terme rare dans une collection documentaire est plus utile à la représentation d'un document le contenant qu'un terme plus fréquent. L'auteure soutient son hypothèse en invoquant qu'un terme représente plus significativement l'information qu'il participe à construire dans un texte s'il est spécifique au sujet abordé et qu'à l'inverse les termes plus génériques sont moins informatifs. Or ce degré de spécificité peut être quantifié dans une collection en observant la fréquence documentaire des termes : moins un terme apparaît dans les documents d'une collection, plus il est considéré comme spécifique à un sujet donné. Parallèlement, un terme figurant dans beaucoup de documents différents est considéré comme plus générique. Afin de représenter les documents de façon à surpondérer les termes spécifiques, plus discriminants, (Sparck Jones, 1972) propose une fonction de pondération reprise plus tard sous le nom d'*IDF* pour *Inverse Document Frequency*. Cette fonction considère comme unique variable la fréquence documentaire d'un terme dans une collection. En en prenant l'inverse, le poids accordé à un terme sera d'autant plus grand qu'il sera rare dans l'ensemble des documents. L'influence positive de cette fonction dans les résultats de RI a été démontrée dans différents travaux précurseurs (Sparck Jones, 1973 ; Salton et C.-S. Yang, 1973) qui ont comparé les performances de systèmes intégrant différentes représentations documentaires. Toutefois, les conclusions présentées tendent à nuancer l'impact de l'*IDF*, en soulignant que son apport est fortement dépendant de la collection documentaire considérée.

En associant cet *IDF* au *TF* précédemment évoqué dans une fonction de pondération globale, le poids accordé à un terme relativement à un document reflète à la fois son degré de spécificité au sein de la collection et son degré de représentativité au sein du document considéré. Le produit de ces deux variables, noté *TF*IDF*, est devenu le schéma

de pondération classique en représentation documentaire, que ce soit pour des tâches de RI, de classification, de *clustering* ou encore de résumé automatique. Plusieurs variantes en ont été proposées, dans le but de normaliser l'une ou l'autre des deux variables, en considérant des paramètres supplémentaires. C'est le cas de la fonction BM25 présentée dans ce qui suit.

Selon la fonction de pondération *TF*IDF*, le poids accordé à un terme t dans un document d est défini telle que dans l'équation (1.1), où *tf(t,d)* correspond à la fréquence d'occurrence de t dans d ; *df(t)* à la fréquence documentaire de t dans la collection documentaire considérée, *i.e.* le nombre de documents dans lequel t apparaît ; et N le nombre total de documents dans la collection.

$$TF*IDF(t,d) = tf(t,d) * log(\frac{N}{df(t)}) \tag{1.1}$$

BM25

Bien que le *TF*IDF* soit régulièrement considéré comme la fonction standard pour la pondération des termes d'indexation en RI, le fait de ne pas considérer la taille des documents dans son calcul lui est souvent reproché. Il est vrai que considérer la fréquence d'occurrence brute comme l'un des ses piliers, alors que la taille des différents documents d'une collection peut grandement varier, intègre un fort biais dans l'attribution des poids : si un document donné *D1*, composé de 100 termes, contenait 10 fois un terme t, ce terme aurait un plus fort poids dans l'index de ce document *D1* que dans l'index d'un document *D2* qui ne contiendrait t qu'une seule fois mais qui ne serait long que de 10 termes. Or en considérant la fréquence relative du terme dans chacun des documents, t a la même distribution dans *D1* et dans *D2*.

Le *TF*IDF* n'est réellement pertinent que si l'on suppose au départ que tous les documents d'une collection ont la même taille. Or c'est assez rarement, voire jamais, le cas dans quelque corpus que ce soit. (Sparck Jones, Walker et al., 2000) proposent alors d'adapter le schéma initial en y intégrant un facteur de normalisation relatif à la taille des documents, en supposant que la différence de longueur entre deux documents relatant un même sujet relève seulement de la verbosité du plus long. Les auteurs s'interrogent par ailleurs sur la façon de mesurer la longueur d'un document : faut-il en compter les caractères ou plutôt les termes ? Si l'on opte pour les termes, mieux vaut-il considérer les types ou les tokens ? Dans leur version brute ou normalisée ? Est-il pertinent de filtrer les mots-vides ou de les considérer dans le décompte ? Cette discussion se conclue sur le fait que l'unité considérée importe peu du moment que tous les documents sont mesurés avec la même, mesure qu'il est par ailleurs pertinent de normaliser en considérant dans le calcul la longueur moyenne des documents de la collection.

L'équation (1.2) présente le calcul du poids BM25 d'un terme t dans un document d, avec les mêmes variables *tf(t,d)*, *df(t)* et N que dans le calcul du *TF*IDF* (1.1). S'y ajoutent dl_d, longueur du document d ; dl_{avg}, longueur moyenne des documents de la collection considérée et deux constantes b et k_1 dépendantes de la collection. Le paramètre k_1 détermine l'impact de l'augmentation du *TF* sur le poids final. Les expérimentations menées par les auteurs sur une des collections test de TREC (T741000X) ont permis de déterminer que la valeur idéale pour cette constante se trouvait dans l'intervalle [1.2,2]. Quant au paramètre b, il s'agit d'un facteur de normalisation dont la valeur doit être comprise dans l'intervalle [0,1]. Les expériences comparatives menées sur la même collection ont déterminées que la valeur optimale de b dans ce cadre était 0.75. Ces valeurs ne peuvent être déterminées, pour une collection donnée, que si celle-ci est annotée, *i.e.* pour laquelle on connaît les documents pertinents et non pertinents relativement aux requêtes considérées. Ces informations n'étant pas toujours disponibles sur des corpus d'expérimentations en RI, ou en faible quantité, les valeurs proposées par (Sparck Jones, Walker et al., 2000) sont régulièrement considérées par défaut dans les implémentations de ce schéma de pondération.

$$BM25(t,d) = \frac{tf(t,d) * (k_1 + 1)}{tf(t,d) + k_1 * (1 - b + b * \frac{dl_d}{dl_{avg}})} * log \frac{N - df(t) + 0.5}{df(t) + 0.5} \tag{1.2}$$

1.3 Mesure de similarité des contenus

Au cœur du processus de recherche d'information se trouve le modèle calculant la similarité entre la requête soumise par l'utilisateur et les documents de la collection considérée. Notons ici que cette requête doit être exprimée dans des termes figurant dans le vocabulaire d'indexation, sans quoi ils ne seront pas considérés dans le processus de RI. Grossièrement, la RI débute par l'expression d'une requête par l'utilisateur, censée rendre compte de son besoin d'information. L'étape suivante compare chacun des termes de cette requête à l'index inversé, contenant l'ensemble des termes d'indexation : si aucun des termes de la requête n'est dans ce fichier-ci, le résultat renvoyé sera nul car tous les documents de la collection sont représentés par ces termes particuliers. Dans le cas du Web, l'index inversé est immense et contient potentiellement l'ensemble des termes possibles d'une langue. En revanche, pour une recherche documentaire centrée sur une collection spécialisée, le vocabulaire d'indexation est régulièrement contrôlé et l'utilisateur à l'initiative d'une requête doit avoir une bonne connaissance du vocabulaire d'indexation pour interroger le système. C'est pourquoi cette tâche était à l'époque l'affaire de spécialistes, auxquels des utilisateurs externes exprimaient leur besoin d'information en langue naturelle, besoin ensuite traduit en requête par les experts.

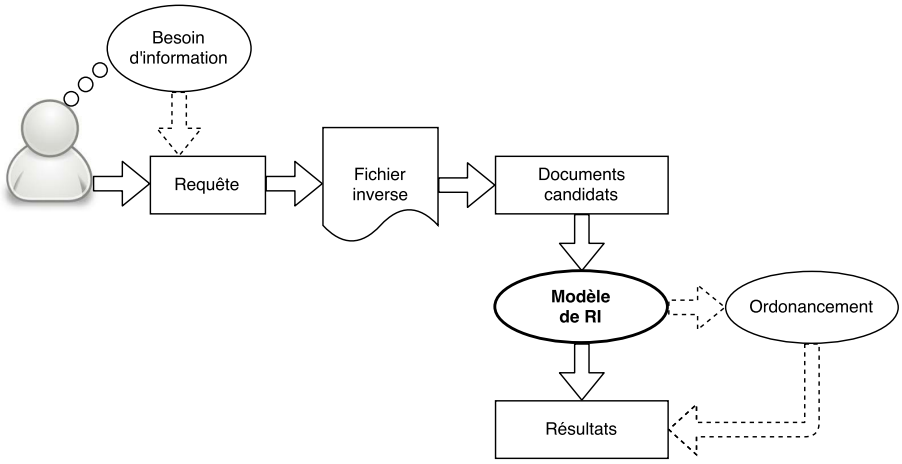

FIGURE 1.1 – Schéma d'un processus de RI

La comparaison de la requête au fichier inverse permet de récupérer un ensemble de documents candidats, *i.e.* globalement ceux partageant au moins un terme de la requête. Comme l'illustre la Figure 1.1, vient ensuite le modèle de RI, consistant à comparer la requête à l'ensemble de ces candidats pour ne conserver finalement qu'un sous-ensemble de documents considérés par le système comme des réponses pertinentes à cette requête. En fonction de ces modèles, les résultats retournés à l'utilisateur peuvent être, ou non, triés par ordre décroissant de pertinence. En d'autres termes, certains modèles se contentent de présenter tous les documents retenus, tandis que d'autres plus précis implémentent une fonction de score. Cette fonction attribue à chacun des documents un score de pertinence qui permet de trier les résultats. Sans aspirer à une présentation exhaustive de l'actuel état de l'art en RI, cette section en détaille certains des modèles les plus connus et les plus implémentés.

1.3.1 Modèle booléen

Le modèle booléen est le plus ancien et le plus simple des modèles de RI. Il repose sur la théorie des ensembles et l'algèbre de Boole en faisant appel aux opérateurs logiques ET, OU et SAUF. Les requêtes soumises au système implémentant un tel modèle prennent elles aussi la forme d'expressions booléennes. En voici une illustration relative à la *collection-exemple* : "*financement* ET *campagne* SAUF *bygmalion*". Le système interroge alors le fichier inverse présenté en Table 1.8 pour récupérer les documents contenant à la fois *financement* et *campagne* mais pas *bygmalion*. Il s'agit donc de trouver l'intersection des ensembles de documents contenant *financement* et *campagne*, puis d'exclure de l'ensemble obtenu ceux contenant *bygmalion*. La Figure 1.2 illustre ce processus de sélection documentaire : dans ces conditions, seul le **Doc 4** répond correctement à la requête soumise.

Ce modèle booléen, relativement intuitif voire naïf dans son concept, présente de nom-

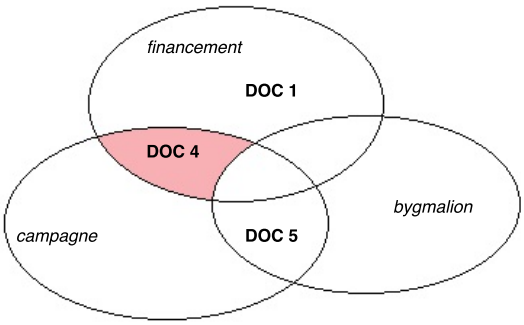

FIGURE 1.2 – Résultats de la requête *financement* ET *campagne* SAUF *bygmalion* sur la *collection-exemple*

breux inconvénients dans le processus de RI. Tout d'abord, la taille de l'ensemble de résultats fournis par un système implémentant un tel modèle peut varier d'un extrême à l'autre. En effet, si la requête soumise correspond à l'union de termes plutôt génériques – ou si l'un d'eux seulement l'est – il en résultera un grand nombre de documents pas nécessairement pertinents en réponse à la requête. À l'inverse, si les termes de la requête sont trop spécifiques, l'ensemble de résultats sera petit, voire vide si la requête constitue une union de termes rares qu'aucun document de la collection ne contient ensemble. Une seconde faiblesse du modèle tient à son fondement théorique, celui des ensembles, qui par définition ne sont pas ordonnés. Du fait par ailleurs que les termes ne soient que binairement pondérés, il est également impossible de trier les résultats sur la base de l'importance des termes qu'ils contiennent. Finalement, les résultats proposés à un utilisateur en réponse à une requête lui sont présentés sans aucun tri, tâche à laquelle il doit lui-même s'adonner. Ces contraintes font du modèle booléen un modèle très strict dont l'absence de souplesse prive l'utilisateur de résultats qui auraient été susceptibles de l'intéresser au regard de la requête soumise. Considérons par exemple une requête représentant l'union de trois termes A, B et C : un document de la collection contenant les termes A et B, mais pas le terme C, sera considéré comme nul en réponse à la requête, au même titre qu'un autre document ne contenant aucun de ces trois termes. Or il aurait pu être pertinent de hiérarchiser les résultats de sorte à pouvoir en proposer certains ne correspondent pas strictement aux contraintes booléennes fixées par la requête soumise.

1.3.2 Modèle vectoriel

Afin de pallier les lacunes du modèle booléen, Gerard Salton propose dans ses travaux (SALTON, 1968 ; SALTON, 1971 ; SALTON, WONG et al., 1975) un modèle plus avancé basé sur une représentation vectorielle des contenus. L'idée est de représenter requête et documents dans un même espace où chacune des dimensions correspond à un terme dis-

tinct présent dans l'ensemble de documents considérés. Cet espace de termes peut s'avérer grand si la collection considérée contient beaucoup de documents et donc beaucoup de termes différents. Le poids d'un terme t dans un document D est représenté par la valeur associée au vecteur de D sur la dimension correspondant à t. Il peut s'agir d'une simple pondération binaire, accordant une valeur de 1 ou de 0 en fonction de l'absence ou de la présence de t dans D, ou d'une pondération plus sophistiquée telles que celles présentées en Section . La Figure 1.3 illustre cette représentation, avec un espace à trois dimensions *t1*, *t2* et *t3*, sur lequel sont projetés les vecteurs de trois documents *D1*, *D2* et *D3*.

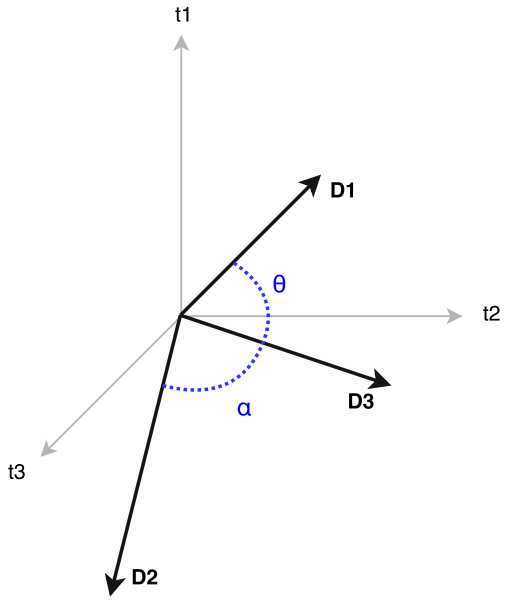

FIGURE 1.3 – Représentation vectorielle dans un espace de termes à 3 dimensions

À partir de cette représentation, il est possible de mesurer la distance entre deux documents, ou bien entre une requête et un document, en mesurant le cosinus de l'angle formé par leurs représentations vectorielles. L'intérêt de ce type de modèle par rapport au modèle booléen tient au fait que la requête puisse être représentée dans le même espace que les documents susceptibles d'y répondre. Ainsi, plutôt que de simplement conserver les documents répondant strictement à ses contraintes, le modèle vectoriel attribue un score à chacun des documents de l'espace, permettant d'ordonnancer les résultats par ordre de pertinence relativement à la requête soumise. Il est alors plus simple pour l'utilisateur à l'initiative de la requête de trouver l'information qu'il recherche, en débutant son parcours documentaire par le document ayant obtenu le plus haut score de pertinence.

L'équation 1.3 présente la formule mesurant la similarité entre une requête q et un document d, considérant à la fois le produit scalaire des deux vecteurs, ainsi que leurs normes respectives. Lorsque les poids affectés aux termes d'index sont positifs ou nuls, le cosinus varie entre 0 et 1. Ainsi, un cosinus élevé rend compte d'un angle faible entre le vecteur requête et le vecteur d'un document donné, révélant une forte similarité entre

les deux. En effet, deux vecteurs identiques auront un angle nul, pour lequel la valeur du cosinus atteindra la valeur maximale de 1. À l'inverse, le cosinus de l'angle formé par deux vecteurs représentant des contenus (requête ou document) ne partageant aucun terme sera de 0, révélant une similarité nulle. Notons également que ce modèle permet non seulement de comparer une requête aux documents d'une collection, mais aussi de comparer différents documents entre eux puisque tous sont représentés dans le même espace vectoriel : chaque paire de documents forment alors un angle dans cette modélisation.

$$sim(q,d) = cos\ \theta = \frac{\vec{q}.\vec{d}}{||\vec{q}||\ ||\vec{d}||} \tag{1.3}$$

Voici un exemple d'utilisation du modèle vectoriel, dans le contexte de notre *collection-exemple*. Soit une requête classique à base de mots-clés Q = "*financement campagne*". L'ensemble de documents candidats résultant de la comparaison à l'index inversé contient les documents **Doc 1, Doc 4** et **Doc 5**. L'espace vectoriel déduit de cette sélection comprend 14 dimensions, correspondant aux 14 termes distincts présents dans les documents candidats, soit l'ensemble suivant :

[*affaire, buisson, bygmalion, campagne, copé, document, financement, lybien, mauvais, mentionner, nouveau, passe, révélation, sarkozy*]

En appliquant une simple pondération binaire des termes d'index, on obtient, pour chacun des trois documents, les scores de cosinus suivants :

- $sim(q,Doc\ 1) = cos(\vec{q}, \vec{Doc1}) = \frac{1}{\sqrt{2}\sqrt{7}} = 0.27$
- $sim(q,Doc\ 4) = cos(\vec{q}, \vec{Doc4}) = \frac{2}{\sqrt{2}\sqrt{7}} = 0.53$
- $sim(q,Doc\ 5) = cos(\vec{q}, \vec{Doc5}) = \frac{1}{2\sqrt{2}} = 0.36$

Contrairement au modèle booléen, chacun des trois documents obtient ici un score, permettant de les ordonner du plus pertinent (**Doc 4**) au moins pertinent (**Doc 1**). Ce modèle vectoriel est donc bien moins restrictif puisqu'il permet à des documents qui ne contiennent pas tous les termes de la requête de figurer parmi les résultats. Il est par ailleurs intéressant de constater que le **Doc 5** obtient un meilleur score que le **Doc 1**, malgré le fait que chacun d'eux ne partagent qu'un unique terme avec la requête, respectivement *campagne* et *financement*. Ceci tient au fait que le **Doc 5** est plus court, et contient donc parallèlement moins de termes non communs avec la requête. En effet, l'une des spécificités de ce modèle tient au fait que sont considérés dans le calcul à la fois les termes communs à la requête et au document considéré et les termes présents dans seulement l'un des deux. L'idée sous-jacente à ce principe relève de la spécificité des documents : si deux documents partagent avec une requête le même nombre de termes mais que l'un présente parallèlement un plus grand nombre de termes non-partagés avec celle-ci, cela signifie qu'il est moins spécifique au sujet décrit par la requête que l'autre

document. Or il est plus pertinent de présenter à l'utilisateur un document répondant spécifiquement à son besoin d'information, traduit en requête, qu'un document y répondant plus génériquement. Notons en revanche que ce modèle ne permet pas de prendre en compte un terme de la requête si celui-ci n'apparaît dans aucun des documents de la collection considérée. L'espace des termes est en effet généré sur la base de l'index inversé, recensant uniquement les termes des documents de la collection, en amont du traitement d'une quelconque requête. Un nouveau terme ne correspondra donc à aucune des dimensions de cet espace et sera ignoré lors de la projection de la requête dans l'espace, ainsi donc que dans le calcul du cosinus.

1.3.3 Modèles probabilistes

Un dernier type de modèle de RI base sa théorie sur une probabilité conditionnelle, celle qu'un document d soit pertinent[7] sachant une requête q, que l'on note $P(R|d,q)$. Le principe fondamental de ce type de modèle est celui de l'ordonnancement probabiliste (noté *PRP* pour *Probability Ranking Principle*) que l'on doit à (Robertson, 1977). Par ce principe, la similarité entre un document et une requête se calcule en comparant les probabilités de pertinence – $P(R|d)$ – et de non pertinence – $P(\overline{R}|d)$ – d'un document d pour une requête q, tel qu'exprimé dans l'équation 1.4. Nous en présentons dans cette section deux modèles parmi les plus populaires, que sont le modèle Okapi-BM25, basé sur la pondération BM25 précédemment présentée, et les modèles de langue.

$$sim(d,q) = \frac{P(R|d)}{P(\overline{R}|d)} \tag{1.4}$$

Lorsque l'on parle du modèle BM25, ou régulièrement Okapi-BM25 en référence au nom du premier système l'implémentant (Robertson et Walker, 1994), on réfère plus au modèle de RI qu'à la fonction de pondération des termes d'index présentée en Section 1.3.2. Le fait est que le modèle est basé sur cette fonction mais qu'il est tout à fait possible de pondérer les termes via la fonction BM25, sans toutefois implémenter un modèle probabiliste. La probabilité de pertinence d'un document d en réponse à une requête q est calculée dans ce modèle par la somme des poids de pertinence de chacun des termes présents à la fois dans d et q. Pour chacun des termes t communs à une requête q et un document d, on calcule donc sa probabilité de pertinence via la fonction $BM25(t,d)$ présentée en équation (1.2). Partant de cette fonction, le calcul de similarité entre q et d

7. La pertinence est régulièrement notée R, pour *Relevance*.

selon ce modèle se présente comme dans l'équation 1.5.

$$\textit{Okapi-BM25}(d,q) = \sum_{t \in q \cap d} BM25(t,d) \tag{1.5}$$

Les modèles de langue sont des concepts basés quant à eux sur les régularités de la langue qu'ils parviennent à estimer *via* des fonctions probabilistes. À partir d'un corpus dans une langue donnée, un modèle de langue va associer à chaque mot (ou séquence de mots) s composant ce corpus une probabilité $P(s)$ grâce à une fonction de probabilité P. Il devient ensuite possible, grâce à cette fonction, d'estimer la probabilité de n'importe quelle séquence de mots de la langue, ou dit autrement, d'estimer la probabilité de *générer* la séquence de mots à partir du modèle calculé (Boughanem et al., 2004).

Appliqué à la RI, cette approche par modèles de langue considère chaque document de la collection comme un sous-langage, pour lequel on calcule un modèle de langue. Une requête q soumise au système est alors considérée comme une séquence de mots, et le score de chacun des documents d correspond à la probabilité que son modèle $M(d)$ génère la requête q. Formellement, la similarité entre une requête q et un document d selon ce type de modèle correspond au produit des probabilités de chacun des termes t_i de la requête, tel que présenté dans l'équation (1.6).

$$P(q|d) = \prod_{i=1}^{m} P(t_i|d) \tag{1.6}$$

Parce qu'il s'agit d'un produit, les documents ne contenant pas exactement tous les termes de la requête se verront assignés une probabilité nulle. Or nous notions plus haut, au sujet des lacunes du modèle booléen, qu'un document pouvait s'avérer pertinent en réponse à une requête sans toutefois en contenir tous les termes. Il est donc d'usage de pondérer les probabilités conditionnelles des termes de la requête $P(t_i|d)$ par un facteur normalisateur, afin d'en calculer une estimation. Différentes méthodes existent en ce sens, dont trois des plus connues sont les celles de Jelinek-Mercer, de Dirichlet et du décompte absolu (*Absolute Discounting*). Nous n'exposons pas ici le détail de ces différentes propositions de lissage de probabilités puisque les modèles de langues, bien que considérés comme performants pour les tâches de RI, ne sont pas implémentés dans ces travaux de thèse. Toutefois, le lecteur souhaitant en apprendre d'avantage sur le sujet pourra se référer à (Zhai et al., 2001), où sont exposées et comparées ces différentes techniques.

1.4 Recherche d'informations d'actualité

Dans le cadre de cette thèse, nous nous intéressons à des données particulières que sont les informations d'actualité, qui connaissent un intérêt grandissant dans la communauté RI depuis les années 1990. Face à la multiplication des sources et des applications de diffusion d'informations, la nécessité d'automatiser la gestion de ces contenus est devenue un enjeu de taille. Parmi les premières propositions dans le domaine, on peut citer les travaux de (Sanderson et al., 1991) ou encore ceux de (Ariki et al., 1997) et (Abberley et al., 1999). Les premiers proposent un SRI interrogeant une collection d'articles de presse issus de différents journaux anglais tels que le *Financial Times* : l'implémentation est simple et se base sur un modèle faisant la somme des poids des termes de la requête pour chaque document de la collection [8]. Les seconds proposent quant à eux un modèle de classification automatique de reportages de chaînes infos, basé sur des mots-clés repérés automatiquement dans un flux de parole puis extraits pour constituer l'index des documents. Les derniers présentent un système d'indexation automatique d'archives de la BBC et un SRI probabiliste. Le système estime des modèles de langue basés sur les trigrammes d'un large corpus de données d'actualité en anglais pour représenter les textes.

En 2013, le programme TREC propose une tâche de résumé temporel d'événements d'actualité (*Temporal Summarization* [9]) en partant du constat que les approches classiques de RI ne permettent pas de traiter efficacement ce type de données fortement marquées temporellement (Aslam et al., 2014). En effet, lorsqu'un nouvel événement tel qu'une catastrophe naturelle ou un accident grave survient et s'inscrit dans l'actualité, très peu d'articles sont immédiatement publiés sur le sujet, car il faut aux journalistes le temps d'investiguer et d'écrire. Toutefois, les internautes, notamment ceux directement impactés par l'événement, ont un besoin urgent d'informations à son sujet, qui ne seront disponibles que quelques heures après. En outre, ces informations délivrées tardivement peuvent s'avérer déjà obsolètes voire fausses, notamment dans le cas d'importantes crises lors desquelles un flot continu d'informations puis de démentis inonde les canaux de diffusion. Il est donc nécessaire de considérer cet aspect évolutif de l'événement d'actualité, et ce dans un temps court. En ce sens, la tâche proposée à TREC consiste à développer des systèmes capables de contrôler et de délivrer des informations au sujet d'un événement d'actualité en temps réel, en diffusant des mises à jour fiables et pertinentes sous la forme de phrases courtes et claires. Les deux éditions suivantes de TREC, de 2014 et 2015, ont de nouveau proposé cette tâche aux participants (Aslam et al., 2015 ; Aslam et al., 2016). Depuis 2016, elle a évolué en intégrant un nouveau type de données qui va croissant, celles issues des réseaux sociaux (Lin et al., 2016).

8. Ce modèle est en fait proche du modèle Okapi-BM25, mais le poids associé à chaque terme n'est ici calculé que par l'IDF du terme dans la collection.
9. http://www.trec-ts.org/

Une importante conférence lié au domaine de la RI (ECIR) a également proposé en 2016 un nouvel atelier de recherche dédié aux problématiques liées aux contenus d'actualité, NewsIR [10] (Martinez et al., 2016). Parmi les trois thématiques proposées alors, deux ont retenu notre attention dans le cadre de cette thèse : celle relative à l'analyse des données d'actualités au sein d'immenses corpus qu'elles constituent ; et celle relative aux événements d'actualités, qu'il s'agit de détecter, de synthétiser et de filtrer automatiquement.

Parmi les travaux présentés lors de cet atelier, on retiendra notamment ceux de (Corney et al., 2016) qui s'intéressent à l'analyse du large corpus d'actualités construit et proposé à l'ensemble des participants : *The Signal Media One-Million News Articles Dataset* [11] (Signal-1M). Les auteurs proposent entre autre de calculer la similarité entre toutes les paires d'articles composant le corpus afin d'en mesurer la proportion de doublons, tâche proche de la nôtre bien que l'objectif diffère. À cette fin, les auteurs optent pour un classique modèle vectoriel afin de mesurer la similarité entre les articles, et démontrent ainsi que chaque article compte en moyenne 2.2 duplicatas. Les auteurs de (Kutuzov et al., 2016) proposent quant à eux une approche pour détecter les événements d'actualités dans les textes. Leur méthode se base sur des plongements lexicaux dont les modèle sont initialement appris à partir d'un large corpus puis quotidiennement mis à jour. Ils peuvent ainsi suivre l'évolution de certains termes, notamment d'entités nommées, et déduire de l'activité de ces termes les tendances dans le flux d'actualités.

L'agence de presse Thomson Reuters propose une méthode pour regrouper des articles relatant un même fait particulier d'un événement d'actualité (Conrad et al., 2016). Leur méthode est basée sur la comparaison et l'agrégation d'articles ou de *clusters* similaires. Ils comparent les articles/clusters entre eux en mesurant deux scores de similarité, l'un basé sur le texte non structuré et l'autre sur les entités nommées et les syntagmes thématiques [12] des textes repérés automatiquement via leur outil propriétaire OpenCalais [13]. Deux espaces sont alors générés, et les vecteurs représentant chacun des types de données y sont projetés respectivement. L'estimation du degré de similarité entre les vecteurs est basée sur un seuil de score fixé empiriquement (différent pour chacun des espaces) : si le score de similarité entre deux articles excède ce seuil, alors ils sont regroupés dans un même *cluster*. Cette méthode nous intéresse particulièrement puisqu'elle considère différents types de termes d'indexation pour représenter les textes. Toutefois, si les auteurs ont ici choisi de construire différents vecteurs pour chaque type de données, nous proposons dans nos travaux un unique vecteur au sein duquel on adapte les pondérations des termes en fonction de leurs caractéristiques.

10. http://research.signalmedia.co/newsir16/
11. http://research.signalmedia.co/newsir16/signal-dataset.html
12. Les auteurs décrivent ces données comme suit : *"domain independent topical phrases"*.
13. http://www.opencalais.com,2016.

1.5 Évaluation des systèmes de RI

Traditionnellement, deux grandes familles de procédés se distinguent, ou peuvent se compléter, pour l'évaluation des résultats d'un SRI : les évaluation orientées *système*, et les évaluations orientées *utilisateur*. L'objectif d'un tel système étant de satisfaire le besoin d'information d'un utilisateur, il est clairement plus pertinent de mettre en place pour en évaluer les performances une méthode orientée utilisateur, mesurant directement la satisfaction de celui-ci quant aux réponses proposées par le système. Si une telle évaluation se fait *a posteriori* du traitement de la requête (*i.e.* l'utilisateur est en charge du jugement de pertinence de chacun des documents retournés par le système), les méthodes orientées système sont au contraire basées sur un jugement de pertinence en amont du traitement de toute requête. Ces méthodes sont généralement préférées à celles orientées utilisateur qui sont plus coûteuses, à la fois en temps et en argent et plus compliquées à mettre en place. Le principe des méthodes orientées système est de soumettre à un panel d'experts un ensemble de requêtes pour lesquelles ils sont en charge d'attribuer un jugement de pertinence relativement à chacun des documents de la collection considérée. Les données ainsi récoltées constituent ce qu'on appelle un ensemble de test qui sert de référence à laquelle comparer les sorties des systèmes automatiques. Pour chaque document proposé par le système en réponse à une requête, on vérifie le jugement de pertinence que l'expert a attribué à ce document pour cette requête : s'il est positif, la réponse du système est considérée comme bonne ; dans le cas contraire, elle est considérée comme mauvaise. Différentes métriques classiques, que nous décriront en détails plus loin dans cette section, rendent compte des performances globales d'un système en combinant les réponses positives et négatives du système pour chacune des paires requête-document possibles.

La procédure d'évaluation précédemment décrite correspond au paradigme de Cranfield qui reproduit en conditions expérimentales un processus de RI (Cleverdon, 1967). Les performances d'un système y sont évalués par les métriques désormais classiques dans le domaine de *précision*, *rappel* et *F-mesure*. La précision mesure le taux de documents pertinents retournés par le système en réponse à une requête, par rapport à l'ensemble des documents retournés. De façon complémentaire, le rappel mesure, parmi l'ensemble des documents pertinents pour une requête, le taux que le système a su retrouver. Ces deux métriques se basent sur les fréquences de vrais positifs (TP), vrais négatifs (TN), faux positifs (FP) et faux négatifs (FN) tels que présentés en Table 1.9 et se calculent telles que présentées en équation 1.7 pour la précision, en 1.8 pour le rappel.

La F-mesure (*cf.* équation 1.9) est quant à elle une simple moyenne harmonique des métriques de précision et rappel, rendant compte de performances plus globales. Un système idéal devrait dans les fait optimiser à la fois la précision et le rappel, mais les valeurs de ces mesures pour un système donné varient régulièrement en sens contraires puisque

SYSTEME \ REFERENCE	PERTINENT	NON-PERTINENT
PERTINENT	TP	FP
NON-PERTINENT	FN	TN

TABLE 1.9 – Matrice de confusion pour une évaluation des performances d'un SRI

lorsqu'une a tendance à augmenter, l'autre a parallèlement tendance à baisser.

Une autre métrique, basée sur ces mêmes paramètres, permet de mesurer le taux de bonnes décisions prises par le système par rapport à l'ensemble des décisions prises. Il s'agit de la mesure d'*exactitude* (*cf.* équation 1.10) qui considère également dans son calcul les documents qu'à la fois l'utilisateur et le système ont jugés non pertinents en réponse à une requête.

$$Precision = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{1.7}$$

$$Rappel = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{1.8}$$

$$F\text{-}mesure = \frac{2 * Precision * Rappel}{Precision + Rappel} \tag{1.9}$$

$$Exactitude = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{1.10}$$

Si ce paradigme reste encore aujourd'hui un procédé incontournable d'évaluation en RI, il présente de nombreuses lacunes régulièrement pointées par les chercheurs de la communauté, dues aux contraintes fortes qu'il impose théoriquement. L'une des plus notables est le fait de considérer les résultats du système comme un ensemble non-ordonné, alors qu'en situation réelle de recherche d'information, il est évident que tous les documents n'ont pas le même pouvoir informatif relativement à la requête soumise. De ce fait, il est important qu'un système soit en mesure de présenter à l'utilisateur les documents les plus pertinents en tête des résultats puis de dérouler les suivants par ordre décroisant de score de pertinence.

Différentes métriques plus récentes, intégrant cette notion d'ordonnancement, ont vu le jour par la suite. On pourra se reporter à (MANNING et al., 2008) pour une présentation détaillée de ces métriques, mais pour n'en citer que quelques exemples aujourd'hui très connus et implémentés, on retiendra notamment :

— La précision à N documents (notée $P@N$) est utilisée dans les cas où le calcul

du rappel présente peu d'intérêt, voire n'est pas possible du tout. Un exemple flagrant est le Web lui-même, dans le cadre duquel il est impossible de connaître le nombre de documents pertinents de la collection qui n'ont pas été retournés par le système. Dans ce cas, les performances du système sont simplement mesurées en observant le nombre de documents pertinents proposés jusqu'à un rang donné, en décomptant par exemple dans le cadre du web combien de bons résultats sont présentés dans la première page, soit environ 10 documents. On cherche alors ici à calculer la précision à 10 documens, notée *P@10*. L'équation ci-dessous (1.11) présente le calcul simple de cette métrique décomptant la fréquence de documents pertinents (r) parmi les N premiers documents retournés.

$$P@N = \frac{r}{N} \tag{1.11}$$

— La précision moyenne, dite *AP* pour *Average Precision*, correspond à la moyenne des valeurs de précision obtenues pour une requête q à différents rangs k, en fonction du rappel. Pour chaque valeur de rappel comprise dans l'intervalle [0,1], on calcule la précision du système pour une requête telle que présentée en équation (1.12), où k correspond au rang considéré ; n au nombre de documents retournés par le système ; *P(k)* à la précision au rang k ; *rel(k)* à une fonction indicatrice dont la valeur est 1 si le document est pertinent pour la requête q, 0 sinon ; N_{rel} au nombre total de documents pertinents pour q. La moyenne des précisions à chacun des rangs est finalement calculée.

$$AP(q) = \frac{\sum_{k=1}^{n}(P(k) * rel(k))}{N_{rel}} \tag{1.12}$$

— La *MAP* (pour *Mean Average Precision*) fait la moyenne des *AP(q)* calculées pour chacune des requêtes q d'un ensemble de test (*cf.* équation 1.13). Elle offre ainsi un aperçu des performances globales du système en considérant simultanément tous les résultats obtenus pour chacun des besoins d'information considérés.

$$MAP = \frac{\sum_{q=1}^{Q} AP(q)}{Q} \tag{1.13}$$

— Le DCG (pour *Discounted Cumulative Gain*), contrairement à toutes les autres métriques précédemment citées, considère une échelle de jugement de pertinence non binaire, plus large. L'idée est que toutes les erreurs du système n'ont pas le même impact et ne doivent pas toutes être pondérées de la même manière. Pour chacun des documents dans un rang donné, le DCG calcule le degré d'utilité – ou de gain – de ce document par rapport à la requête et en fait finalement la somme. Les deux variables permettant ce calcul sont d'une part le rang occupé par le document (i dans l'équation 1.14) et d'autre part la valeur du jugement de pertinence qui

lui a été attribué (rel_i dans l'équation 1.14). Notons que sur une échelle de valeurs initialement donnée, le jugement de pertinence est proportionnel à la satisfaction du juge par rapport au résultat (*i.e.* plus la valeur est élevée, plus le document est pertinent). Cette méthode permet donc de privilégier les systèmes capables de présenter en tête de résultats les meilleurs documents, pour ne présenter les moins bons qu'en fin de parcours. La valeur obtenue par un SRI selon cette métrique reflète donc sa capacité à ordonnancer correctement les résultats qu'il sélectionne comme pertinents.

$$DCG = \sum_{i=1}^{p} \frac{rel_i}{log_2(i+1)} \tag{1.14}$$

Un autre biais du paradigme de Cranfield est sa difficulté d'implémentation dans le cadre de grandes, voire très grandes, collections documentaires. Dans son article sur la philosophie de l'évaluation en RI, (Voorhees, 2001) remet en question l'exhaustivité des jugements de pertinence dans de tels contextes. En effet, il est difficilement envisageable de demander aux experts d'attribuer un jugement de pertinence à chacun des documents, pour chacune des requêtes de l'ensemble de test, si la collection documentaire compte plusieurs centaines de milliers de documents. Dans l'hypothèse où un juge nécessiterait 30 secondes de réflexion pour l'attribution d'un jugement de pertinence, il faudrait au total 30*800 000 = 24 400 000 secondes = 277 jours = 9 mois pour juger chacun des documents d'une collection de 800 000 entrées, et cela pour une seule requête.

Afin de surmonter cette incapacité, (Sparck Jones et Van Rijsbergen, 1975) proposaient dès les années 70 de ne considérer qu'un sous-ensemble de la collection documentaire dans la tâche d'attribution de jugements de pertinence. Cette méthode, dénotée plus tard par le terme de *pooling*, consiste à soumettre l'ensemble des requêtes considérées dans l'ensemble de test à l'ensemble des systèmes que l'on souhaite évaluer et comparer. Chacun des systèmes proposent en sortie un ensemble de résultats, parmi lesquels on ne retient que les N premiers – au minimum 50 pour que l'ensemble de test final ne soit pas trop biaisé, d'après (Sparck Jones et Van Rijsbergen, 1976). Les résultats respectifs de chacun des systèmes sont réunis dans un ensemble final qui constitue le sous-ensemble de documents à juger relativement à la requête initiale. Si cette méthode permet d'approcher la représentativité d'un ensemble de test, l'incomplétude qui le caractérise a été de nombreuses fois remise en question. Le choix de N, nombre de documents retournés par le système conservés pour l'évaluation, est notamment discuté dans (Buckley et al., 2007). Par ailleurs, il est peu recommandé, dans un contexte de *pooling*, de tester les performances d'un nouveau système, non considéré lors de la génération du sous-ensemble de documents évalués. En effet, un système radicalement différent de ceux initialement testés serait susceptible de remonter de nouveaux résultats pertinents, non jugés lors de la phase

d'évaluation. Or dans ce contexte de *pooling*, tout document non jugé par un expert est par défaut considéré comme non pertinent pour une requête. Les résultats obtenus pour ce nouveaux système ne seraient donc pas représentatifs de ses performances réelles.

Qu'ils soient exhaustifs ou non, ces ensembles de test sont très coûteux à construire. Ils nécessitent un important investissement humain et par conséquent, d'importants moyens financiers pour rémunérer les juges en charge de l'évaluation. Si certains grands industriels peuvent se permettre une telle dépense pour récolter les données nécessaires à l'évaluation puis à l'évolution de leurs systèmes propriétaires, les chercheurs de la communauté ont plus de difficultés à mettre en place ce type de procédure. Par ailleurs, si l'objectif d'une équipe de recherche est de comparer ses propres résultats à ceux d'autres équipes, il est essentiel que tous puissent évaluer les systèmes qu'ils développent sur les mêmes données, sans quoi aucune conclusion ne serait possible.

Partant de cette nécessité, de nombreuses collections de test de grande ampleur ont été développées dans le cadre de conférence internationales. TREC, pour *Text REtrieval Conference*, est à ce jour certainement l'une des des plus réputées. Depuis 1992, soutenu par le NIST (*National Institue of Standard and Technologies*), ce cycle de conférences a proposé à la communauté scientifique de nombreuses tâches à résoudre dans le domaine de la RI, en fournissant d'importants ensembles de test dédiés à l'évaluation des systèmes en compétition. Le site internet de TREC [14] offre en accès libre l'intégralité des résultats proposés lors des différentes éditions, ainsi que les ensembles de test qui sont accessibles et donc exploitables dans des contextes extérieurs aux conférences. Des collections de test sur des tâches plus spécialisées sont également disponibles, telle que celles du CLEF [15] (*Cross-Language Evaluation Forum*), en place depuis les années 2000 et essentiellement axé sur la RI dans un contexte multilingue et/ou multimodal. On compte également parmi les plus importantes les collections proposées par le projet NTCIR [16] (*NII Testbeds and Community for Information access Research*), qui existe depuis 1997.

Par ailleurs, sans développer de collections de test spécifiques, il existe d'autres conférences particulièrement notables dans le domaine de la RI. SIGIR [17] (*Special Interest Group in Information Retrieval*) et son équivalent Européen ECIR [18] (*European Conference on Information Retrieval*), tout deux débutés en 1978, sont encore à ce jour des rendez-vous incontournables pour les acteurs du domaine. Plus récemment en 2004, les rencontres CORIA [19] (Conférences en Recherche d'Information et Applications) ont vu le jour en France, tout en revendiquant une ouverture internationale.

14. http://trec.nist.gov/
15. http://www.clef-initiative.eu/
16. http://research.nii.ac.jp/ntcir/index-en.html
17. http://sigir.org/
18. La conférence n'a pas de site propre, mais chacun des éditions annuelles en dispose d'un, dont le dernier en date est accessible via ce lien : http://ecir2017.org/
19. https://asso-aria.org/index.phpoption=com_content&view=article&id=72&Itemid=474

1.6 Conclusion

Nous avons présenté dans ce chapitre un état de l'art du domaine de la recherche d'information. Loin d'être complet au regard des évolutions permanentes du domaine depuis sa naissance dans les années 1950 à ce jour, auxquelles un manuscrit entier aurait pu être consacré, nous nous sommes attachés à décrire les étapes au cœur du processus d'automatisation que cette tâche représente.

Nous avons notamment exposé les difficultés que posait la représentation des contenus textuels, qui ne sont pas structurés logiquement de façon à ce qu'un système automatique puisse en saisir clairement le sens. Les réponses apportées à ces difficultés relèvent de techniques de segmentation et de normalisation textuelles, ainsi que de méthodes d'extraction d'information, grâce auxquelles la phase d'indexation des contenus aboutit à des représentations formelles qu'un système automatique est en mesure d'appréhender.

Nous sommes par la suite revenus sur les grandes familles de modèles de recherche d'information, que sont les approches booléennes, vectorielles et probabilistes. Si les plus récentes d'entre elles, les méthodes à base de probabilités, sont à ce jour considérées comme les plus performantes, le modèle vectoriel avait avant elles connu un succès conséquent, en dépassant l'approche booléenne initiale, généralement reconnue comme trop restreinte pour modéliser la tâche cognitive complexe qu'est la RI.

La dernière partie de ce chapitre était dédiée à une présentation succincte des méthodes d'évaluation en RI, qui constituent encore aujourd'hui un défi majeur pour les chercheurs de la communauté. Pour témoigner d'une évolution scientifique dans quelque domaine que ce soit, par rapport à l'existant, il est essentiel de pouvoir en opérer une comparaison objective. Dans le domaine de la RI, un des critères fondamentaux pour approcher cette objectivité est de comparer différents systèmes sur les mêmes données et c'est de ce constat que se développent depuis les années 1990 de larges collections de test. Ces questions au sujet de l'évaluation nous ont particulièrement occupés dans ce travail de thèse et nous reviendrons plus amplement sur celles-ci dans la seconde partie de ce manuscrit décrivant nos contributions.

Chapitre 2

Topic Detection and Tracking

Le domaine du *Topic Detection and Tracking* (TDT) est né d'un programme de recherche américain sponsorisé à la fin des années 90 par le DARPA [1], qui souhaitait pouvoir traiter rapidement et efficacement des flux d'informations d'actualité en les regroupant pertinemment en fonction des sujets abordés. Le rapport résultant de cette étude pilote (Allan, Papka et al., 1998) constitue aujourd'hui la référence incontournable à tous travaux relevant de cet objet d'étude.

Si la tâche est régulièrement dénotée en français par l'expression "*détection et suivi d'événement*" (Binsztok et al., 2002 ; Bossard et al., 2008), nous préférons la traduire par "*détection et suivi de sujets à caractère événementiel*", qui présente clairement la distinction entre les notions de *sujet* et d'*événement*, sur lesquelles nous revenons dans ce chapitre.

Nous commençons en Section 2.1 par présenter la tâche et ses objectifs, ainsi que les différentes sous-tâches qui la composent. Nous revenons ensuite en Section 2.2 sur l'une de ces sous-tâches, la *Story Link Detection*, dans laquelle s'inscrit particulièrement certaines composantes de la thèse. La Section 2.3 expose par la suite les considérations temporelles intervenant dans le traitement des données spécifiques que sont les informations d'actualité. Nous revenons enfin, avant de conclure ce chapitre, sur les campagnes et méthodes d'évaluation traditionnelles de cette tâche, en Section 2.4.

2.1 Présentation de la tâche

Le cadre du TDT décrit dans (Allan, 2002) définit un sujet (en anglais, *topic*) comme un ensemble de contenus d'actualité (en anglais, *news stories*) cohérent dans le fait qu'ils aient en commun un événement majeur (en anglais, *trigger event*) fonctionnant comme déclencheur du sujet qu'ils relatent. Un événement est quant à lui défini dans (Y. Yang, J. G. Carbonell et al., 1999) comme "*quelque chose de non trivial survenant à un certain endroit à un certain moment*" [2], à l'origine d'un nouveau sujet. Par exemple, les

1. *Defense Advanced Research Projects Agency*
2. *"Something non-trivial happening in a certain place at a certain time".*

attentats de Paris de la nuit du 13 novembre 2015 ont été l'événement déclencheur d'un sujet dont on parle encore aujourd'hui, au travers d'articles relatant une arrestation, une cérémonie d'hommage ou un témoignage de victime. Tous ces articles, relatant des faits qui n'existent que parce qu'est survenu cet événement majeur, sont considérés comme faisant partie d'un même ensemble, d'une même histoire, d'un même sujet.

Plus précisément, un sujet n'est ici pas considéré dans le sens commun qu'on lui confère intuitivement, celui qui caractériserait la réponse à la question "*De quoi parle cet article ?*". L'auteur de (Allan, 2002) fait une distinction entre *topic* et *subject*, que le français permet difficilement de discerner. Un *subject* est défini comme une notion plus large que le *topic* dont il est une sous-catégorie caractérisée par un ancrage temporel dont d'autres formes de *subjects* ne sont pas empreintes. Il est en effet possible d'observer dans le flux quotidien d'informations des nouvelles aux sujets atemporels tels que *la conquête de l'espace* ou *la recherche contre le cancer*, qui ne sont pas initiées par un événement originel. Le critère définitoire pour faire d'un *subject* un *topic* a donc trait, dans ce cadre tout du moins, à l'existence d'un événement déclencheur de sujet. C'est cet objet d'étude précis qui intéresse le domaine du TDT, d'où notre proposition d'en traduire le nom par l'expression précise et complète de "*détection et suivi de sujets à caractère événementiel*". Notons que la distinction entre *topic* et *subject* peut être ambiguë du fait qu'un événement à l'origine d'un *topic* peut lui-même constituer un *subject* : en reprenant le précédent exemple, la série d'attaques de Paris constitue l'élément déclencheur du *topic*, mais lorsqu'il est relaté dans un article, sa mention permet aisément de répondre à la question "*De quoi parle cet article ?*". Afin de lever toute ambiguïté terminologique, nous emploierons par la suite l'anglicisme *topic* pour désigner l'objet d'étude du champ de recherche décrit dans ce chapitre.

Concernant la notion d'événement, la définition précédemment exposée, bien que sommaire, la délimite assez nettement. Elle considère, en mentionnant son caractère *non-trivial*, un événement comme un fait qui rompt l'ordinaire, qui survient sans que l'on s'y attende. S'ajoutent à ce critère des paramètres d'ancrage spatio-temporels, définissant un événement comme un fait circonscrit à un lieu précis, dans un intervalle de temps relativement court (même s'il arrive qu'un événement s'éternise, il est régulièrement ponctuel). Ainsi, une attaque terroriste quelconque, si elle rompt bien l'ordinaire, ne constitue pas en soi un événement. En revanche, celle de Paris le 13 novembre 2015, ou celle d'Ankara le 10 octobre 2015 constituent bien deux événements distincts du même type.

La dernière notion sur laquelle nous revenons est celle de *news stories*, qui recouvre dans le cadre du TDT à la fois des dépêches d'actualité et des transcriptions d'émission télévisées, radiophoniques et Web. Cette dimension multimédia des contenus traités constituaient, à l'époque de l'émergence du domaine, une avancée par rapport aux champs de recherche classiques s'intéressant aux contenus de presse, qui limitaient leur objet d'étude

au texte. Dans la suite de ce chapitre, nous dénoterons cette notion par le terme de *nouvelle*, qui nous semble refléter le sens initial tout en considérant l'aspect multimédia qui la caractérise dans ce contexte.

C'est devant le constat d'un flux massif et incessant de contenus d'information que l'initiative du projet TDT s'est concrétisée. Il était devenu nécessaire de gérer efficacement ces données, d'être capable de les *ranger* pour en simplifier l'analyse et y repérer rapidement des informations importantes. L'objectif initialement formulé était de pouvoir segmenter un texte pour en isoler les différentes nouvelles abordées, afin de regrouper entre elles celles relatant un même *topic*, et parallèlement de détecter celles ne relatant aucun *topic* connu, pour lesquelles il s'agissait d'en considérer un nouveau.

Afin de résoudre cette tâche, le premier rapport (Allan, Papka et al., 1998) proposait une segmentation en trois sous-tâches, qui se sont par la suite, dans des travaux plus récents (Allan, 2002), subdivisées en cinq sous-tâches que sont :

— *Story Segmentation*
La segmentation de nouvelles a pour but de découper en nouvelles individuelles un contenu textuel en relatant plusieurs. Cette tâche concerne essentiellement les transcriptions d'émissions qui, à l'image d'un journal télévisé, abordent différentes nouvelles qu'il convient d'identifier.

— *First Story Detection*
La détection de la première mention d'un *topic* consiste à repérer, dans le flux d'information, un *topic* inédit, jamais détecté auparavant. En d'autres termes, il s'agit de pouvoir décider qu'une nouvelle n'appartient à aucun des *topics* précédemment considérés comme tels.

— *Cluster Dectection*
La détection de groupes est une extension de la tâche précédente, en ce sens qu'elle consiste à regrouper dans différents groupes toutes les nouvelles référant à un même *topic*. Mais lorsqu'une nouvelle constitue une première mention de *topic*, alors un nouveau groupe doit être créé. Il s'agit d'une tâche non supervisée, pour laquelle le nombre de groupes n'est pas connu par le système, qui doit lui-même le déterminer au regard des données qui lui sont fournies.

— *Topic Tracking*
Le suivi d'actualité nécessite la surveillance continue du flux d'informations pour en extraire des actualités relatives à un *topic* particulier que l'on spécifie au système. Cela s'apparente à une tâche de recherche et de filtrage d'information, pour lequel on fournit en entrée du système un ensemble de contenus dont on sait qu'ils réfèrent à un même *topic*, et qui doit fournir en sortie tous les nouveaux contenus du flux relatifs à ce *topic*.

— *Story Link Detection*

> La détection de lien entre nouvelles peut être considérée comme une méta-tâche, utile à la résolution de toutes les autres. En effet, son objectif est de déterminer si deux nouvelles que l'on soumet au système relèvent ou non du même *topic*. Il s'agit donc de savoir représenter les nouvelles de manière à pouvoir les comparer, puis de déterminer une fonction de similarité rendant compte de leur degré de proximité thématique. Un système robuste capable de résoudre cette tâche serait en mesure d'influer positivement sur la résolution des trois précédentes, pour lesquelles la comparaison de nouvelles est fondamentale.

C'est cette dernière tâche de *Story Link Detection* qui nous intéresse particulièrement au regard des problématiques qui sont les nôtres dans cette thèse. En tant que méta-tâche, elle soulève des questions ayant trait à la représentation des données dans ce contexte spécifique, ainsi qu'aux méthodes de calcul de similarité entre contenus d'information, auxquelles nous souhaitons également répondre. La différence par rapport à nos propres travaux est que ce qui est ici soumis au système est une paire de nouvelles, pour lesquelles il s'agit de décider si elles relatent le même *topic*. Dans notre contexte, même si l'on peut grossièrement reconsidérer la chose ainsi, on ne peut envisager de comparer un article soumis au système à toutes les vidéos composant la collection documentaire. Les méthodes proposées pour la résolution de cette tâche ne constituent donc qu'une partie de ce que nous souhaitons réaliser, une brique au cœur du système développé. La section suivante en propose un bref état de l'art, présentant les principales méthodes implémentées pour répondre à ces problématiques.

2.2 Story Link Detection

Cette sous-tâche peut être considérée comme une composante des autres sous-tâches du TDT : un système la traitant constitue une brique essentielle pour la résolution de chacune des autres. Pour cette raison, peu de travaux se sont attachés à en proposer un système de résolution intrinsèque, privilégiant ceux de résolution des autres tâches, intégrant notamment cette composante. Néanmoins, la littérature du domaine présente certains travaux en proposant des modèles de résolution, développant principalement diverses méthodes de représentation des nouvelles, et diverses mesures saisissant la similarité entre deux nouvelles.

En 1999, une équipe participe à la campagne d'évaluation TDT199 (*cf.* Section 2.4) et propose un article (Brown et al., 1999) comparant deux modèles de résolution de cette tâche, plus tard repris et synthétisé dans (Y. Yang, J. Carbonell et al., 2002). Pour chacune des méthodes proposées, une représentation vectorielle des documents basée sur

une pondération TF*IDF est implémentée pour modéliser les contenus [3]. Mais si la première méthode, notée CMU-1 [4], intègre la fréquence absolue d'un terme dans un document comme valeur de TF, la seconde méthode (CMU-2) intègre le logarithme de cette valeur, notée log(TF). Par exemple, pour un document contenant trois fois le terme x, on a pour cette combinaison document-terme $TF = 3$. CMU-1 attribue donc pour le vecteur de ce document une valeur de 3 pour la dimension x. En revanche, CMU-2 attribue au vecteur de ce même document la valeur de $1+log(3) = 1.48$ pour cette même dimension x. Par ailleurs, les données du corpus utilisé pour le développement des systèmes sont divisées en deux parties : l'une pour l'entraînement des modèles, l'autre, plus petite, réservée à la phase de test. Alors que CMU-1 calcule les poids TF*IDF relativement à l'ensemble des données, CMU-2 se contente de ceux de la base de test pour calculer les siens. C'est cette fois donc l'IDF des termes qui se trouve modifié d'une version à l'autre. Nous reprenons l'exemple du terme x, dont on suppose qu'il apparaît dans 100 documents d'un corpus initial en comptant 1000. En divisant le corpus en deux sous-ensembles, on suppose que l'ensemble de test de 200 documents en comprend 12 contenant le terme x. l'IDF de x relativement à l'ensemble du corpus, implémenté dans CMU-1, est alors de $log(1000/100)=1$, tandis que l'IDF implémenté dans CMU-2 est de $log(200/12)=1.22$. Cette différence influence de nouveau la représentation vectorielle des contenus pour chacune des méthodes.

La manière de comparer les vecteurs est en revanche identique pour les deux versions CMU-1 et CMU-2. Il s'agit dans les deux cas de mesurer le cosinus de l'angle formé entre le vecteur requête et chacun des vecteurs candidats pour obtenir une estimation de leur degré de similarité respectif. Nous retrouvons donc ici une implémentation classique du modèle vectoriel décrit dans (Salton, Wong et al., 1975), présenté au Chapitre 1.

Au regard du protocole d'évaluation décrit plus bas en Section 2.4, les résultats de (Brown et al., 1999) démontrent de notables variations de performances entre CMU-1 et CMU-2, qui ne se distinguent pourtant que par de simples adaptations de coefficient de pondération. Si CMU-1 présente globalement de meilleures performances sur l'ensemble des données du corpus, CMU-2 est largement meilleur lorsque l'on compare leurs performances respectives obtenues sur les données de l'ensemble de test. Ces résultats tendent à démontrer l'importance de la capacité de généralisation des systèmes, qui doivent être capable de s'adapter aux nouvelles données, et donc pouvoir s'affranchir de connaissances qu'ils peuvent en avoir *a priori*. Par ailleurs, ces résultats témoignent de l'importance à attacher à la représentation des contenus textuels dans ce type de tâche, qu'il ne faut pas négliger au profit de l'optimisation des mesures de similarité entre documents.

En ce sens, (Tsagkias et al., 2011) propose de comparer différentes représentations

3. Le chapitre 1 présente en détails les métriques et modèles de RI exposés dans cette section.
4. L'équipe de recherche proposant ces méthodes est affiliée à Carnegie Mellon University, dite CMU, d'où l'identifiant associé aux différents systèmes proposés.

documentaires pour résoudre une tâche d'appariement d'informations en ligne et de médias sociaux. Ils se basent sur la structure particulière des contenus de presse en ligne, dont on peut distinguer différentes sections telles que le titre, le chapô, le corps de l'article, ainsi que les métadonnées sur l'auteur et la source. Ils supposent également que des informations sémantiques telles que des entités nommées peuvent être extraites afin de modéliser plus précisément les contenus décrivant des événements, incluant généralement des acteurs, dénotés par des noms de personnes ou d'organisations, et des lieux. Différentes représentations intégrant ces différents types d'informations sont comparées, et il ressort des résultats présentés que les meilleures performances sont atteintes en considérant les représentations *sac de mots* considérant à la fois le titre et le corps de l'article. Notons par ailleurs que le modèle de similarité choisi par (Tsagkias et al., 2011) dans ces travaux est un modèle de langue, calculant pour une paire *(R,D)* la probabilité que le document D génère la requête R, au regard des termes qui les composent respectivement.

Lors de la campagne d'évaluation TDT2002, l'ensemble des équipes engagées dans la résolution de cette tâche ont choisi un modèle vectoriel à base de similarité cosinus couplée à une pondération TF*IDF des termes pour représenter les documents. Le système présenté par l'équipe de l'université du Massachusetts, décrit dans (Allan, Lavrenko et al., 2002), propose les meilleurs résultats pour cette tâche. Ils sont régulièrement considérés comme *baseline* pour les travaux plus récents.

En 2004, cette domination du modèle vectoriel est remise en question par les auteurs de (F. Chen et al., 2004) qui proposent une comparaison de plusieurs méthodes alternatives à celle-ci. Ainsi, ils proposent d'y opposer trois modèles basés respectivement sur la distance de Hellinger, l'indice de Tanimoto (Duda et al., 1973) et la mesure de clarté (Croft et al., 2001). Le premier correspond à une mesure de probabilité entre la distribution des termes dans deux documents que l'on cherche à comparer ; le second décrit un rapport entre le nombre de termes communs à deux documents, et le nombre de termes présents dans un seul des deux ; et enfin le dernier est un modèle de langue se basant sur la comparaison des distributions de probabilités des termes dans un document donné et en anglais général. Implémentés séparément, il apparaît que le modèle de langue produit le système offrant les meilleures performances, immédiatement suivi par le modèle vectoriel. En revanche, en comparant différents systèmes implémentant différentes combinaisons de scores, les meilleures performances sont atteintes par celui incluant l'ensemble des quatre mesures précédemment citées. Ces résultats témoignent de l'intérêt de prendre en compte des indices de différentes nature dans le calcul du score final, pour le rendre plus objectif en s'affranchissant des spécificités d'un modèle unique.

Il ressort de ce bref état de l'art que la tâche de *Strory Link Detection* a unanimement été reconsidérée comme une tâche de recherche d'information, dont les méthodes sont exploitées pour sa résolution. Que ce soit par le biais de modèles vectoriels ou de modèles

de langue, la comparaison binaire de nouvelles se fait toujours par un calcul de similarité entre deux représentations documentaires, dont on cherche à saisir le degré de proximité. Sans proposer de nouvelles méthodes de calcul, les travaux dans ce domaine se sont en revanche attachés à rechercher et comparer diverses modélisations des contenus dont l'importance dans la résolution de la tâche s'est avérée essentielle.

2.3 Considérations temporelles

Lorsqu'il est donné à traiter de l'information en ligne, et notamment des données d'actualités tel que l'objet d'étude du TDT, la prise en compte de la dimension temporelle est incontournable. (Y. Yang, J. G. Carbonell et al., 1999) écrivent en effet que deux nouvelles décrivant le même événement ont tendance à être proches temporellement, ce qui suggère de se baser sur des critères à la fois sémantique et temporel pour leur comparaison.

Le rapport de l'étude pilote présente les propositions de deux participants à la tâche de détection de nouveau *topic* dans un flux d'actualités (Allan, Papka et al., 1998). Les deux équipes développent pour résoudre cette tâche une méthode de *clustering*, dans laquelle une nouvelle est soumise au système qui doit déterminer si elle est à intégrer à l'un des *topics* existants, ou s'il faut crée un *cluster* supplémentaire définissant un nouveau *topic*. L'équipe de l'université Carnegie Mellon (Brown et al., 1999) intègre le paramètre de fraîcheur des résultats en introduisant un filtre temporel dans la sélection des *topics* possible pour une nouvelle N : ne sont considérés comme candidats que les *topics* dont la date de production de la dernière nouvelle ajoutée n'excède pas un nombre de jours donné par rapport à la date de la nouvelle N à catégoriser. L'équipe de l'université du Massachusetts (Allan, Papka et al., 1998) propose quant à elle une adaptation du seuil de score au-delà duquel le système considère qu'une nouvelle appartient à un *topic*, en fonction du temps. Lorsqu'elle arrive, une nouvelle est comparée à chacun des *topics* existant, et ne sont retenus que ceux dont le score de similarité excède un certain seuil initialement fixé. Ce seuil s'adapte à chaque *topic* candidat, et augmente en fonction de la distance temporelle séparant la nouvelle à catégoriser de la dernière nouvelle ajoutée au *topic* : en d'autres termes, le score de similarité entre une nouvelle est un *topic* doit être d'autant plus haut que la distance en nombre de jours est élevée, pour considérer que cette nouvelle relate ce *topic*. En privilégiant ainsi la fraîcheur du résultat, les participants démontrent qu'un événement à tendance à être relaté dans un cours intervalle de temps, puis qu'il est ensuite rapidement délaissé dans le flux d'actualités, au profit de plus récents. Néanmoins, (Brown et al., 1999) concluent sans donner le détails de leurs résultats que l'intégration de critères temporels n'avait au mieux aucune incidence sur les performances, mais pouvait au pire dans certains cas les dégrader. En revanche, (Allan, Papka et al., 1998) démontrent que l'ajout d'une pénalité relative au temps augmente les performances

globales du système.

Si la contrainte temporelle peut être intégrée sous forme de distance constante à ne pas excéder, ou de pénalité sur les scores de similarité, elle peut également être directement intégrée au calcul du score de similarité, tel que décrit dans (Tsagkias et al., 2011). Les auteurs proposent une fonction de similarité globale, considérant des scores représentant différents facteurs, pour la résolution de détection de liens entre contenus d'actualités. Le premier score est un score de similarité thématique tel que décrit en Section 2.2. Le second est un score de date, proposé dans l'idée de privilégier un contenu récent pour une nouvelle soumise en requête, sans toutefois exclure les contenus plus anciens : la première étape du calcul est une probabilité conditionnelle basée sur un intervalle temporel t, qui vaut 0 si la date de production d'un résultat candidat est hors de cet intervalle, et une valeur inversement proportionnelle au nombre de nouvelles publiées dans cet intervalle sinon. La seconde étape est un lissage appliqué dans le but de réduire les écarts entre les scores des candidats récents, dont la date est incluse dans t, et des plus anciens, dont la date est hors de t. Le score final est finalement calculé par une simple moyenne géométrique de chacun des scores, *i.e.* en prenant la racine *n-ième* du produit des n scores à considérer : en l'occurrence, les auteurs de (Tsagkias et al., 2011) ajoutent un troisième score spécifique aux données qu'ils traitent et qui nous intéresse peu, et implémentent donc en fin de traitement la racine cubique du produit des trois scores calculés. Ils démontrent qu'en comparaison à un modèle équivalent n'intégrant aucun paramètre temporel, leur système présente de meilleures performances en introduisant un score de fraîcheur au calcul du score global décrivant le lien entre deux nouvelles.

La problématique sous-jacente est en fait de savoir comment modéliser le compromis à trouver entre similarité des contenus et fraîcheur du résultat dans cette tâche de détection de lien entre nouvelles sensibles à ces deux paramètres. Par ailleurs, certains systèmes de traitement de contenus d'information dépassent le cadre du TDT, et traitent un plus large éventail de sujets dont certains ne sont pas temporellement ancrés. Il est alors question de pouvoir déterminer quelles sont, parmi les données à traiter, celles pour qui il convient de considérer la dimension temporelle, des autres pour lesquelles une mesure de similarité de contenu suffit. (Dong et al., 2010) s'intéressent à la problématique de l'ordonnancement des résultats dans le contexte du Web, et débutent par la proposition d'un classifieur distinguant binairement les requêtes en fonction de leur sensibilité temporelle. Leur objectif est d'apprendre un modèle d'ordonnancement des résultats d'une requête, en se basant sur des paires requête-document décrites par différents attributs. Ces attributs dénotent notamment le rang occupé par un document parmi l'ensemble des résultats possibles pour une requête, en fonction de sa pertinence d'une part, et en fonction de sa fraîcheur lorsque la requête le nécessite. Prenons un exemple fictif, pour lequel les juges en charge de l'annotation estiment qu'une requête R a pour ensemble de résultats les documents *{D1, D2,*

D3, D4}. Du fait que ces résultats ne soient pas également pertinents en réponse à R, un premier tri est fait en fonction de leur niveau de cohérence thématique, permettant d'obtenir l'ensemble ordonné *{D4, D1, D3, D2}*. La requête R étant sensible à la fraîcheur du résultat, un second tri est fait indépendamment, sur la base de critères temporels, dont l'ensemble résultant est cette fois *{D2, D1, D4, D3}*. Finalement, deux ensembles de données d'apprentissage sont construits indépendamment. Un premier, que l'on note $C_{General}$, intègre toutes les paires requête-document du corpus, que les requêtes soient ou non sensibles à la fraîcheur du résultat. Ces instances sont décrites par un ensemble d'attributs rendant compte du degré de similarité de leurs contenus. Un second, que l'on note $C_{Fraicheur}$, intègre uniquement les paires requête-document pour lesquelles la requête est sensible à la fraîcheur du résultat, qui sont cette fois décrites par un ensemble d'attributs rendant compte de leur proximité temporelle. Notons que la taille de ce second ensemble est bien plus restreinte que celle du $C_{General}$. Afin de pallier ce sous-effectif dans l'échantillonnage des données d'apprentissage, les auteurs proposent deux méthodes distinctes. Une première consiste à fournir au modèle d'apprentissage les données des deux ensembles, en accordant un poids plus fort à celles issues de $C_{Fraicheur}$. Une seconde est basée sur des méthodes d'adaptation, qui permettent en classification d'apprendre dans un premier temps un modèle sur un ensemble de données d'un domaine (en l'occurrence $C_{General}$), puis de l'adapter aux données d'un autre domaine ($C_{Fraicheur}$) dans un second temps. C'est cette dernière méthode qui permet en moyenne d'obtenir les meilleurs résultats d'ordonnancement considérant à la fois la similarité thématique et la fraîcheur des résultats.

Ce tour d'horizon nous permet de constater qu'en fonction des données et de la problématique que l'on souhaite résoudre, diverses possibilités sont envisageables pour considérer la dimension temporelle dans l'appariement de contenus d'actualité. Il semble en revanche, mis à part dans un cas, qu'un consensus se fasse autour de l'importance de considérer cette dimension dans le traitement d'un type d'actualités très présentes en ligne, celles relatant des sujets *chauds* (Charon, 2010).

2.4 Évaluation

2.4.1 Campagnes et Corpus

Dans le but d'évaluer les performances d'un système de TDT, et dans celui d'en comparer différentes méthodes de résolution, plusieurs corpus de données de référence ont été construits. En se basant sur ces données, les chercheurs investis dans cette tâche ont de quoi estimer les performances de ce qu'ils conçoivent, en comparant les résultats que leurs systèmes prédisent à ceux dont on sait qu'ils sont vrais. Ces différents corpus sont

constitués par des experts dans le cadre de campagnes destinées à l'évaluation de tâches précises. Cette section détaille les évolutions des campagnes proposées pour l'évaluation du TDT, ainsi que des corpus de données constitués pour en rendre compte.

Le premier corpus destiné à l'évaluation des systèmes expérimentaux de TDT décrits dans (Allan, Papka et al., 1998) a été constitué entre 1996 et 1997 par les membres du groupe de projet pilote eux-mêmes, avec le soutien du *Linguistic Data Consortium*[5] (LDC). Ce consortium a soutenu par la suite les développements de tous les corpus dédiés à l'évaluation des tâches relatives au TDT. Dans cette première version, les données constituant l'ensemble correspondent à des dépêches de Reuters (équivalent de notre AFP), ainsi que des transcriptions de reportages de CNN. Au total, 15 683 nouvelles composent ce corpus, au sujet d'actualités s'étant déroulées entre janvier et juin 1995. Parmi elles, 25 *topics* ont été sélectionnés sur la base de leur forte couverture, garantissant ainsi un nombre conséquent de nouvelles du corpus y référant. Un ensemble de juges experts s'est chargé de l'évaluation exhaustive de ces *topics* : chacune des nouvelles du corpus est jugée en fonction de chacun des *topics* retenus. Il s'agit d'un jugement binaire, voué à indiquer pour chaque paire (N, T) si la nouvelle N réfère ou non au *topic* T. Au total, ce sont 392 075 jugements de pertinence qui ont été attribués pour la constitution de ce corpus.

La campagne associée à ce corpus, notée TDT1997, donnait pour objectif aux participants la résolution de quatre tâches fondamentales : la segmentation, le suivi d'actualités, la détection de groupes tels que décrits précédemment en Section 2.1, ainsi que la détection rétrospective. Cette dernière correspond à une variante de détection de groupes pour laquelle les participants connaissent à l'avance le nombre de groupes à constituer. Il s'agit donc cette fois-ci d'une tâche de classification supervisée. Les organisateurs de la campagne d'évaluation souhaitaient ainsi évaluer l'impact de l'absence de connaissance sur les résultats de regroupement. Il est évidemment plus simple pour un système de regrouper correctement les nouvelles par *topic* s'il sait dès le départ le nombre de *topics* à considérer.

La campagne suivante, notée TDT1998, intègre toutes les tâches de la précédente, exceptée celle de détection rétrospective. Elle se base sur un nouveau corpus d'évaluation – TDT-2, (Cieri et al., 1999) — composé d'environ 57 000 nouvelles issues de dépêches et d'émissions de radio et de télévision recueillies entre janvier et juin 1998. Un protocole d'évaluation manuelle plus précis que pour la campagne précédente a été proposé par le LDC, grâce auquel 100 *topics* ont été sélectionnés et annotés par les juges experts, relativement à chacune des nouvelles disponibles.

Une évolution majeure caractérise les campagnes suivantes de 1999 et 2000 (*resp.* TDT1999 et TDT2000), puisqu'elles intègrent dans leurs jeux de données des contenus

5. https://www.ldc.upenn.edu/

en chinois, tandis que les corpus précédents étaient exclusivement en anglais. Ce nouveau corpus – TDT-3, (Graff et al., 1999) — s'est vu ajouté, par rapport aux précédents, des nouvelles de deux autres sources en anglais, ainsi que de trois sources en chinois mandarin, recueillies entre octobre et décembre 1998. Au total, 71 388 nouvelles constituent ce corpus. Ces nouvelles chinoises sont ajoutées sous forme de transcriptions dans leur langue originale, ainsi que dans une version traduite en anglais. Le corpus est alors bien plus conséquent que les précédents, et intègre dans les données une dimension internationale de l'information dont les corpus précédents n'étaient pas empreints. Dans ces deux campagnes, les cinq tâches fondamentales du TDT étaient proposées aux participants, et il était convenu d'utiliser les données du corpus TDT-2 comme données d'apprentissage, afin de réserver celles de TDT-3 pour la phase de test. Pour que les jeux de données des deux corpus soient comparables, et donc qu'un système appris sur l'un puisse s'adapter aux données de l'autre, les organisateurs TDT1999 et TDT2000 ont étendu TDT-2 en y ajoutant des données en chinois mandarin, extraites des trois mêmes sources d'information que TDT-3 et recouvrant des actualités de janvier à juin 1998.

Les deux campagnes suivantes, de 2001 et 2002 (respectivement TDT2001 et TDT2002), sont équivalentes à celles de 2000, tant par les tâches proposées que par les données à traiter. Si les sources utilisées pour générer le corpus TDT-4 sur lequel elles s'appuient sont en effet les mêmes que pour TDT-3, elles reflètent néanmoins une actualité plus proche temporellement de la date des campagnes qui les exploitent, puisqu'elles ont été recueillies entre octobre et janvier 2001. 60 *topics* y ont été aléatoirement sélectionnés et jugés pour constituer l'ensemble de référence.

Les plus récentes campagnes en date sont celles de 2003 et 2004 (notées TDT2003 et TDT2004), qui comme les deux précédentes ont actualisé les données du corpus TDT-3 dans un nouvel ensemble, TDT-5, en récupérant des actualités issues des mêmes sources entre avril et septembre 2003. Par ailleurs, la dimension multilingue du domaine a de nouveau évolué lors de ces sessions, puisque ce sont ajoutées à ces données anglaises et chinoises des données en arabe, recueillies de trois sources différentes sur le même intervalle de temps.

Mise à part la première campagne proposée à la suite de l'étude pilote en 1997, toutes les autres ont été organisées par le *National Institute of Standards and Technologies*[6], sur le site duquel sont accessibles toutes les données succinctement décrites ici, ainsi que les résultats de l'ensemble des participants de chacune des sessions, pour chacune des tâches.

6. NIST:http://www.itl.nist.gov/iad/mig/tests/tdt/

2.4.2 Métriques

Toutes les sous-tâches du TDT sont évaluées en tant que tâche de détection (Fiscus, Doddington et al., 1999 ; Fiscus et Doddington, 2002), par une procédure binaire considérant les sorties du système comme vraies ou fausses par rapport aux sorties attendues.

Alors que les procédures d'évaluation des systèmes de recherche d'information considèrent des mesures de gain telles que la précision et le rappel que les participants cherchent à maximiser, la procédure classique en TDT s'appuie sur des mesures de pénalités, que les participants cherchent à minimiser (Allan, J. G. Carbonell et al., 1998 ; Fiscus, Doddington et al., 1999 ; Allan, 2002). Il s'agit des taux d'omissions et de fausses alertes, complémentaires des mesures de précision et rappel, telles que, en observant la Table 2.1 :

1. Mesures de pénalité :
   — Fausses Alertes = $\frac{FP}{FP+TN}$
   — Omissions = $\frac{FN}{FN+TP}$
2. Mesures de gain :
   — Précision = $\frac{TP}{TP+FP}$
   — Rappel = $\frac{TP}{TP+FN}$

REFERENCE / SYSTEME	Cible	Non-Cible
Cible	TP	FP
Non-Cible	FN	TN

Table 2.1 – Matrice de confusion des résultats d'un système de résolution de TDT

Le choix de considérer d'évaluer les performances du système en fonction de pénalités plutôt que de gains tient au seul fait qu'elles soulignent l'importance de minimiser les erreurs dans la tâche de détection (Wayne, 2000). Par ailleurs, (H.-H. Chen et al., 2002) signalent que ces métriques rendent plus précisément compte de ce qu'un utilisateur observe en situation de recherche d'actualités relatives à un *topic* particulier. Si le système lui propose trop de mauvais résultats (fausses alertes), ou à l'inverse pas assez de bons résultats (omissions), il ne sera pas satisfait. En revanche, le fait de dénombrer les bons résultats parmi l'ensemble proposé (précision) ou parmi l'ensemble des bons qui auraient pu être proposés (rappel), lui importe peu du point de vue de sa démarche de recherche d'information.

Les taux d'omissions et de fausses alertes sont des variables intégrées dans le calcul d'une fonction de coût, que les systèmes cherchent à minimiser pour justifier de bonnes

performances. La fonction de coût pour les tâches de TDT, ramenées à des tâches de détection, est régulièrement notée C_{Det}, telle que :

$$C_{Det} = C_{Omission} * P_{Omission} * P_{Cible} + C_{Fausse-Alerte} * P_{Fausse-Alerte} * P_{Non-cible} \quad (2.1)$$

où :

— $C_{Omission}$ et $C_{False-Alerte}$ correspondent aux coûts respectivement attribués aux cas d'omissions et aux cas de fausses alertes. Il s'agit de paramètres spécifiés *a priori* en fonction de la tâche à accomplir, déterminant le poids à accorder à chaque type d'erreur dans le calcul du coût total. Pour la campagne d'évaluation TDT2 (Cieri et al., 1999), ces deux coûts valent 1 : les deux types d'erreurs sont considérés comme équivalents au regard des tâches à résoudre. En revanche, pour d'autres tâches, ces coûts peuvent se distinguer, comme c'est le cas dans les travaux de (Cselle et al., 2007) qui cherchent à développer un système de suivi d'actualités et de détection de nouveau *topic*, pour lequel ils estiment qu'une erreur d'omission est bien plus grave qu'une erreur de fausse alerte. En d'autres termes, ils préfèrent voir associer à un *topic* des nouvelles qui n'y sont pas liées, plutôt que d'en omettre qui en relèvent effectivement. Le choix est alors fait, pour ces tâches-ci, d'accorder les poids suivants : $C_{Omission} = 1$ et $C_{Fausse-Alerte} = 0.1$.

— $P_{Omission}$ et $P_{Fausse-Alerte}$ décrivent respectivement les probabilités d'omission et de fausse alerte, calculées à partir de la distribution des résultats du système observés en sortie, par rapport aux résultats de référence (tels que décrits précédemment par la Table 2.1).

— P_{Cible} et $P_{Non-cible}$ (*i.e.* *1-*P_{Cible}) sont les probabilités *a priori* qu'une instance appartienne à la classe P_{Cible}, et respectivement qu'une instance n'appartienne pas à la classe P_{Cible}. Ce paramètre dépend de la probabilité de détection pour le corpus analysé, et doit donc être fixé en fonction des données traitées et de la tâche à résoudre relativement à ces données. Plus concrètement, pour une tâche de suivi d'actualités, P_{Cible} correspond à la probabilité qu'une nouvelle appartienne à un *topic* : pour la campagne TDT1998, cette probabilité est calculée sur le corpus TDT2, et obtient la valeur de 0.02. La même valeur est obtenue pour la tâche de détection de *topic*. Pour la tâche de segmentation en revanche, la probabilité P_{Cible} calculée sur ce même corpus atteint la valeur de 0.3.

Cette fonction de coût est normalisée comme suit :

$$C_{Det}(Norm) = \frac{C_{Det}}{MIN(C_{Omission} * P_{Cible}, C_{Fausse-Alerte} * P_{Non-cible})} \quad (2.2)$$

Les probabilités d'omission et de fausse alerte varient régulièrement en sens inverse, quand l'une diminue, l'autre a tendance à augmenter. L'objectif pour un système automatique est donc de trouver le meilleur compromis possible pour minimiser les deux, dans le but final d'abaisser le coût. La tâche étant binaire, il s'agit de trouver un seuil de score départageant les résultats des deux classes 1 et 2, et décider que tous ceux dont le score est supérieur au seuil appartiennent à la classe 1, tandis que les autres appartiennent à la classe 2. Prenons un exemple plus concret avec une tâche de *Stroty Link Detection*, pour laquelle il s'agit de déterminer si deux contenus appartiennent au même *topic* (classe 1) ou non (classe 2). Disons qu'un calcul de cosinus ait été implémenté pour saisir la similarité entre deux contenus : sa valeur pour chacune des paires soumises au système se trouve comprise entre 0 et 1. La question est alors de savoir à partir de quelle valeur de cosinus le système peut considérer que deux contenus relatent le même *topic*, donc de déterminer un seuil de score distinguant les deux classes.

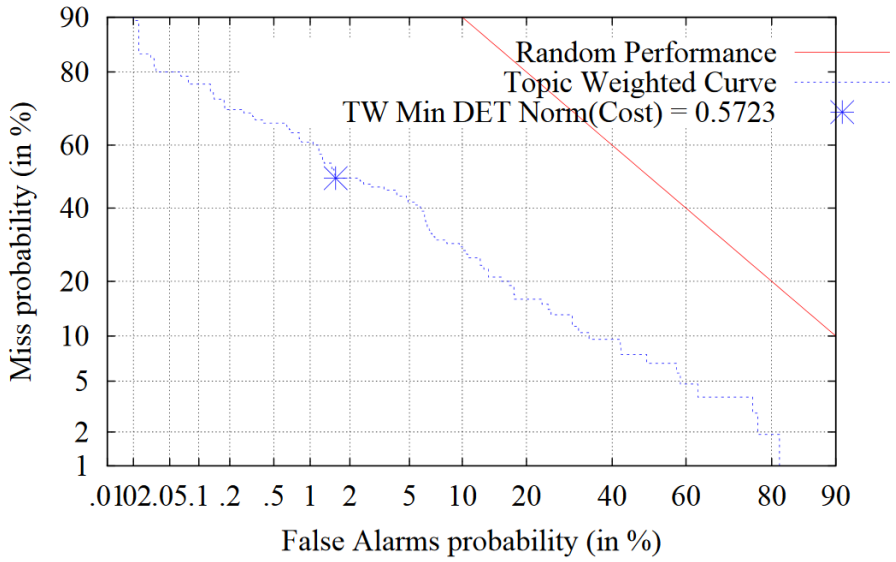

FIGURE 2.1 – Illustration de la courbe *DET*, décrivant les performances d'un système de TDT en fonction des probabilités de fausses alertes et d'omissions pour différentes valeur de seuil entre 0 et 1.

Afin de sélectionner ce seuil de score de similarité optimal, minimisant au mieux les deux types d'erreur, (MARTIN et al., 1997) propose de se baser sur le tracé de la courbe *DET* (*Detection Error Tradeoff*), représentant les points dont les coordonnées décrivent les taux d'omissions et de fausse alertes à différentes valeurs de seuil comprises entre 0 et 1. Pour chacune de ces valeurs de seuil, les probabilités des deux types d'erreurs sont calculées, ainsi que le coût normalisé. Le seuil pour lequel les probabilités d'erreurs procurent le coût minimum est considéré comme optimal.

Une illustration de courbe *DET*, reprise de (KUMARAN et al., 2004) est présentée en Figure 2.1, avec en abscisses la probabilité de fausses alertes et en ordonnées la probabilité

d'omissions. L'étoile bleue représente les probabilités d'omissions et de fausses alertes pour lesquelles la fonction de coût est la plus faible (0.57), et le seuil de score ayant permis au système de produire ces taux d'erreurs est retenu pour le système final.

2.5 Conclusion

Dans ce chapitre, nous sommes revenus sur le domaine du *Topic Detection and Tracking*, autrement dit *détection et suivi de sujets à caractère événementiel*, en en présentant les objectifs et les spécificités.

Nous avons commencé par présenter les notions fondamentales de ce domaine de recherche, dont l'objet d'étude est le *topic*, sujet d'information marqué temporellement, initié par un événement déclencheur, et relaté dans différentes nouvelles d'un flux d'actualités continu.

Par la suite, nous avons rappelé la division de l'objectif en cinq sous-tâches, pour lesquelles il s'agit grossièrement de segmenter un flux d'informations (1) et d'y distinguer les nouvelles relatant un *topic* connu qu'il s'agit de regrouper (2) des nouvelles relatant un *topic* inédit (3) qu'il s'agit d'identifier comme tel. Lorsqu'un *topic* est identifié, il s'agit de suivre le flux d'actualités afin d'en détecter les nouvelles y référant (4) et se baser pour se faire sur des méthodes de comparaison de nouvelles (5). Nous avons insisté sur cette dernière composante, celle du *Story Link Detection*, méta-tâche dont l'objectif est de modéliser la similarité entre deux nouvelles, au cœur de nos problématiques. Reconsidérée comme une tâche de recherche d'information, divers modèles ont été proposés pour résoudre cette tâche, et c'est un modèle vectoriel, majoritairement adopté, que nous avons choisi d'implémenter dans nos propres travaux.

Le chapitre s'achève par la présentation des campagnes et corpus dédiés à l'évaluation des travaux de ce domaine, multilingues et multimodaux, puis de la fonction de coût rendant compte des performances des systèmes qui cherchent à en minimiser la valeur. Les conclusions de cet état de l'art ont servi de base au développement de notre système d'appariement d'articles en ligne et de vidéos, dont la représentation et la comparaison de nouvelles constituent le socle.

Chapitre 3

Recommandation basée sur le contenu

Les systèmes de recommandation trouvent leurs origines dans différents domaines de recherche qui les ont précédés, puisque les premiers d'entre eux sont nés de travaux ayant trait aux sciences cognitives, à la recherche et au filtrage d'information, ou en lien avec des secteurs plus industriels comme le marketing.

Ce n'est qu'au milieu des années 1990 que la tâche de recommandation a réellement émergée en tant que champ de recherche à part entière, avec des travaux précurseurs tels que (Resnick et al., 1994 ; Hill et al., 1995 ; Shardanand et al., 1995). Cette naissance suit de près celle du Web, que l'on situe en 1989, avec lequel la production et la consommation de données, produits et services sont entrées dans une nouvelle dimension qui ne cesse d'évoluer. La recherche d'information permettait d'ores et déjà d'opérer un tri parmi cette masse d'informations, afin de n'en sélectionner que d'infimes parties répondant à des besoins spécifiques. Cependant, cette masse de données est telle qu'il reste souvent compliqué pour les usagers d'y retrouver quoi que ce soit susceptible de les intéresser.

Nous introduisons ce chapitre par une définition de la tâche et de ses objectifs (Section 3.1), en distinguant les deux principales approches proposées dans la littérature. Nous présentons plus en détails en Section 3.2 le modèle générique d'un système de recommandation, ainsi que les variables qu'il implique. Nous nous focalisons ensuite sur les propositions pour la recommandation d'actualités (Section 3.3), et présentons des méthodes de représentation des articles et des utilisateurs puis d'autres pour le filtrage d'informations. Nous discutons finalement en Section 3.4 de la vaste question de l'évaluation de tels systèmes, difficile à appréhender au regard de la complexité des modèles développés pour répondre à cette tâche particulière.

3.1 Présentation de la tâche

L'objectif d'un système de recommandation est de présenter à ses utilisateurs des contenus personnalisés, en fonction de critères qui leur sont propres. Dans ce contexte, les

contenus, qu'il s'agisse d'actualités, de morceaux de musiques ou d'objets divers et variés, sont dénotés sous le terme d'*items*. Quant aux caractéristiques définissant un utilisateur particulier, elles constituent ce qu'il est d'usage d'appeler un *profil utilisateur*. Partant de ces deux types de données, un système de recommandation cherche à proposer à un utilisateur les items de sa base de données les plus pertinents relativement à son profil.

On distingue généralement trois types d'approches pour la construction d'un système de recommandation (Balabanović et al., 1997), dépendant du type d'information considérée (Adomavicius et Tuzhilin, 2005). Les premières sont dites basées sur le contenu (*content-based system*), et considèrent la comparaison des représentations d'items aux profils modélisant les centres d'intérêts respectifs de chacun des utilisateurs. Les secondes sont liées au filtrage collaboratif (*collaborative filtering*), où sont considérés uniquement les profils utilisateurs. L'idée générale de ces approches est que les items sélectionnés par un utilisateur x seront susceptibles d'être pertinents pour un utilisateur y dont les centres d'intérêts sont proches de ceux de x. Les troisièmes méthodes sont dites *hybrides*, puisqu'elles mêlent des méthodes de recommandation basées sur le contenu et des méthodes basées sur le filtrage collaboratif pour améliorer la précision du modèle.

Dans ce chapitre, nous nous intéressons uniquement aux méthodes de recommandation basées sur le contenu, puisque c'est en s'inspirant de celles-ci que le système développé dans le cadre de cette thèse a été reconsidéré, sous certains de ses aspects, comme une tâche de recommandation. Globalement, le profil d'un nouvel utilisateur d'un système de recommandation est initialisé par défaut, puisque le système n'en connaît pas les centres d'intérêts *a priori*. Puis au fil des items que cet utilisateur consulte et note, le système peut déduire quels sont ses centres d'intérêts particuliers, grâce auxquels il sera en mesure de lui proposer par la suite de nouveaux items pertinents relativement à ceux-ci. La recommandation est donc un processus itératif, qui met à jour le profil d'un utilisateur dès que celui-ci émet un jugement de pertinence, implicite ou explicite, au sujet d'un item donné. Les méthodes développées pour modéliser le profil utilisateur en fonction de ses jugements de pertinence nous intéressent particulièrement dans cet état de l'art des systèmes de recommandation. Dans nos propres travaux, nous ne désirons pourtant pas distinguer les utilisateurs entre eux, puisque l'objectif du système est au contraire d'associer une seule et même vidéo à un article, censée convenir à un maximum d'utilisateurs. En revanche, nous souhaiterions pouvoir construire le profil d'un macro-utilisateur, représenté par un échantillon d'usagers du système, dont les avis au sujet des appariements article-vidéo proposés automatiquement guideraient le système pour affiner ses critères de sélection d'une vidéo pour un article.

Par rapport à un système de RI classique, qui cherche à ordonner les résultats retournés par le système, la recommandation relève plus d'une tâche de filtrage d'information, qui cherche à exclure certains contenus de l'ensemble des résultats. Dans ce but, il doit

déterminer un seuil de score, distinguant les items de la collection pertinents pour un utilisateur de ceux qui ne le sont pas, et relève en cela plus d'une tâche de classification binaire.

Nous détaillons dans les sections suivantes les méthodes classiques de représentation des items et des profils utilisateurs dans le cadre de systèmes de recommandation basée sur le contenu. Sont ensuite présentées les méthodes évaluant les résultats de ces systèmes, dont les données recueillies constituent les entrées de modèle d'apprentissage supervisé servant à la mise à jour des profils utilisateurs.

3.2 Modèles de recommandation

Théoriquement, un système de recommandation cherche à estimer, pour un item de la collection, quelle note lui sera attribuée par un utilisateur donné. Pour calculer cette estimation, les modèles se basent sur les notes attribuées par ce même utilisateur à d'autres items précédemment consultés. Le modèle ainsi généré permet d'attribuer des estimations à l'ensemble des items non consultés par l'utilisateur, afin de lui présenter celui ou ceux, en fonction de l'application souhaitée, ayant obtenu(s) la/les plus haute(s) estimations.

Formellement, on considère généralement C un ensemble d'utilisateurs et I un ensemble d'items composant une collection associée à une application donnée. On cherche alors à mesurer l'utilité u d'un item i pour un utilisateur c, via une fonction telle que présentée en (3.1), où E décrit l'espace d'évaluation, c'est-à-dire les jugements possibles que les utilisateurs peuvent attribuer aux items. Il peut s'agir d'un simple jugement binaire du type *j'aime* ou *je n'aime pas*, ou d'une échelle plus large, comme l'attribution d'un nombre d'étoiles reflétant l'intensité de la satisfaction, qui sont les plus répandus sur le Web.

$$u : C \times I \rightarrow E \tag{3.1}$$

À partir de cette définition, l'objectif est de proposer à tout utilisateur $c \in C$ les items $i^* \in I$ qui maximisent sa fonction d'utilité u, tel que décrit en (3.2).

$$\forall c \in C, \quad i^* : arg \max_{i \in I} u(c, i) \tag{3.2}$$

Cette fonction d'utilité ne peut être construite que si le système dispose de données sur ses utilisateurs. Lorsqu'un nouvel utilisateur fait appel à un système de recommandation, il n'en connaît pas les centres d'intérêt, et lui propose donc des contenus aléatoirement, en lui demandant en retour de juger ces contenus sur une échelle de notation graduelle.

À partir de ces jugements de pertinence, le système est en mesure de modéliser le profil de cet utilisateur, auquel seront comparés ensuite des contenus de la base de données qu'il n'a encore jamais consultés. La Figure 3.1 illustre ce processus itératif, dans lequel le profil est mis à jour dynamiquement au rythme des jugements de pertinence exprimés par l'utilisateur.

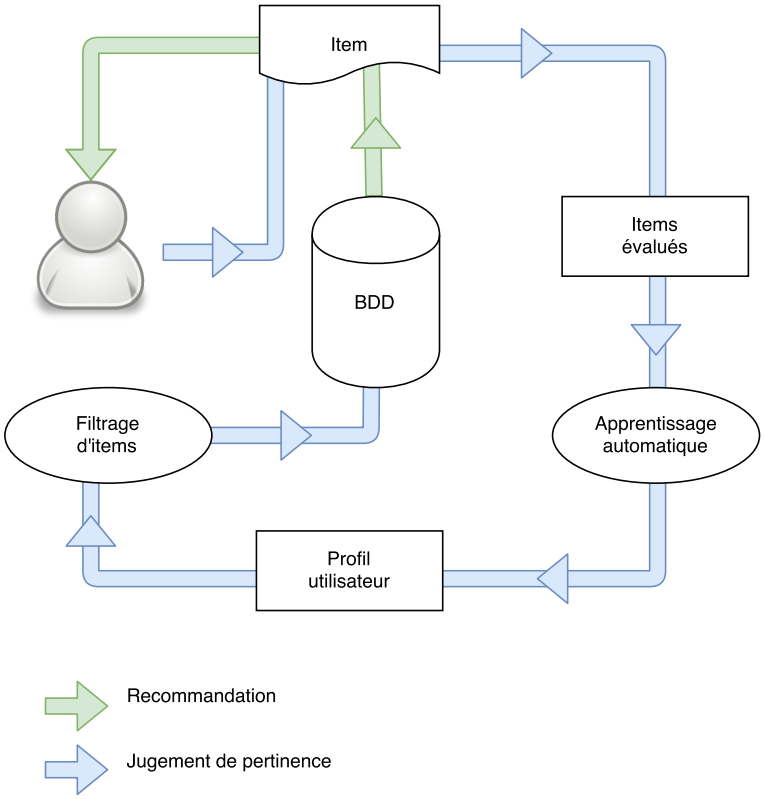

FIGURE 3.1 – Processus itératif d'un système de recommandation

L'explosion du *big data* ces dernières années a vu se développer bon nombre de systèmes de recommandation. Nous nous intéressons ici particulièrement aux moteurs de recommandation de contenus d'actualité, qui traitent des données textuelles similaires aux nôtres. Nous souhaitons notamment observer la façon dont les items, *i.e.* des articles de presse, et les profils utilisateur sont modélisés puis comparés entre eux.

3.3 Recommandation d'actualités

Les premiers systèmes de recommandation d'actualités ont émergé au milieu des années 90, de la nécessité de filtrer ces contenus abondants et incessants pour des utilisateurs submergés par une telle quantité d'information. La plupart des systèmes développés dans ce cadre ont opté pour des approches à base de filtrage collaboratif, ou des méthodes hybrides, mêlant ces dernières à d'autres basées sur du filtrage d'information.

Sans considérer le recours au filtrage collaboratif, qui nous concerne peu dans ces travaux, le processus général d'un système de recommandation d'actualités comprend trois modules que sont :

1. Le module de représentation des articles
2. Le module de représentation des utilisateurs
3. Le module opérant la comparaison des deux représentations et le filtrage des items à soumettre à un utilisateur donné

Nous présentons par la suite quelques uns des travaux relevant de cette tâche, en distinguant les stratégies proposées pour chacun des modules cités.

3.3.1 Représentation des articles

Dans le contexte particulier de la recommandation d'actualités, les items considérés sont des contenus textuels. Comme dans la plupart des systèmes traitant ce type de données non-structurées, les textes sont modélisés sous forme de vecteurs des termes les composant. De la même façon qu'en RI, les acteurs du domaine de la recommandation souhaitent optimiser la représentation vectorielle en sélectionnant au mieux les termes décrivant les items à classer.

L'un des premiers systèmes de recommandation d'actualités, décrit dans (Kamba et al., 1995), récupère quotidiennement les articles de différentes sources du Web et les projette dans un espace de termes en interrogeant le système smart (Salton, 1971) qui implémente la classique pondération *TF*IDF* des termes.

De la même façon, le système NewsWeeder décrit par (Lang, 1995) travaille sur une large collection et conserve l'ensemble des termes extraits de tous les documents qui la composent, moyennant quelques filtres sur les mots-vides. La taille des vecteurs qu'ils traitent varie, en fonction des expérimentations, entre 20 000 et 100 000 dimensions, ce qui correspond aux ordres de grandeur qui sont les nôtres. Les auteurs soulignent malgré tout la difficulté de généraliser un modèle avec autant de dimensions, qui nécessite un très large ensemble de données d'apprentissage pour parvenir à généraliser l'apport d'un terme particulier de ce grand espace. Ils expérimentent une réduction de dimensions par SVD (*Single Value Decomposition*), méthode de factorisation de vastes matrices creuses en matrices denses, mais les résultats de classification obtenus démontrent de moins bonnes performances que celles obtenus sur les données représentées par l'ensemble des dimensions. Une hypothèse pouvant expliquer cette conclusion est, selon eux, qu'une telle réduction d'espace entraîne une perte d'information dans la représentation des contenus.

Syskill & Webert, système décrit dans (Pazzani et al., 1996) et régulièrement repris comme référence, opère quant à lui une sélection des termes représentant les articles.

L'ensemble des termes considérés pour la représentation des contenus est en fait dépendant de l'utilisateur. C'est à partir des données récoltées des annotations de l'utilisateur que le système sélectionne les termes qu'il estime les plus représentatifs de ses préférences. Les nouveaux articles sont ainsi vectorisés et projetés dans cet espace réduit. Un même article obtient donc des représentations différentes en fonction du profil utilisateur auquel il est comparé. Par ailleurs, en réduisant ainsi le nombre de dimensions, les auteurs se content d'implémenter une fonction booléenne pour pondérer les termes de chacun des vecteurs.

De plus récents travaux, comme (Cantador et al., 2008), proposent d'intégrer des informations sémantiques aux documents, tout en conservant une représentation vectorielle. Leur méthode de vectorisation se base sur une interaction avec diverses ontologies de domaine et le Web sémantique, grâce à laquelle les dimensions de l'espace ne représentent cette fois plus de simples termes, mais des concepts plus généraux. Contrairement aux précédents systèmes cités, le vocabulaire utile à la représentation des items est ici complètement contrôlé et structuré, puisque ses composants sont extraits de ressources.

3.3.2 Représentation des profils utilisateurs

Saisir les centres d'intérêt d'un utilisateur nécessite de récupérer en amont certaines informations à son sujet. Celles-ci sont déduites du comportement de l'utilisateur face aux articles qui lui ont été soumis par le système. Ces comportements se distinguent en deux catégories. On parle de retour explicite lorsque l'utilisateur est impliqué dans une tâche de jugement de l'item soumis. Mais il existe en parallèle des indices implicites dénotant l'intérêt de l'utilisateur en réponse à un article proposé. Des critères comme le temps passé sur la page, la vitesse de défilement ou encore l'action d'agrandir certaines zones du texte sont autant d'indices permettant d'estimer l'attention portée par l'utilisateur à un item particulier. Nous reviendrons en Section 3.4 sur ces différentes méthodes d'évaluation des systèmes de recommandation.

L'initialisation du profil utilisateur se fait donc à partir d'un ensemble initial de paires $<I,R>$ où R correspond au jugement de pertinence attribué par l'utilisateur à l'item I.

Une façon simple de représenter le profil de l'utilisateur dans cette tâche particulière est de construire un vecteur de termes dont les poids traduisent ses préférences. C'est ce que proposent les auteurs de (Kamba et al., 1995), qui séparent binairement les articles en distinguant les bons des mauvais sur la base de retours à la fois implicites et explicites. Dans leur modèle, chaque terme de l'espace se voit associer un poids calculé en fonction des jugements attribués aux articles le contenant. Un article jugé positivement augmentera le poids des termes y figurant et, à l'inverse, un article jugé négativement sanctionnera les termes qui le composent. C'est finalement une moyenne de la somme obtenue pour chaque terme qui représente son poids. Ce dernier est par ailleurs nuancé en

fonction de l'*activité du terme* dans les documents consultés par l'utilisateur. Un terme dont le poids est régulièrement mis à jour signifie qu'il est particulièrement présent dans les articles parcourus par l'utilisateur, il est alors considéré par les auteurs comme un indice siginificatif des préférences de celui-ci. Il est en revanche plus compliqué d'estimer l'importance d'un terme peu présent dans les articles jugés : son poids pourrait être faible du seul fait qu'il n'ait pas eu l'occasion d'être augmenté, et non du fait qu'il est déprécié de l'utilisateur.

Une alternative proche de celle-ci est présentée dans (De Gemmis et al., 2015), où les auteurs proposent de distinguer deux vecteurs prototypiques reflétant les préférences utilisateur. L'un correspond aux termes des articles jugés bons, tandis que l'autre représente celui des termes des articles jugés mauvais. La comparaison des scores obtenus pour chaque article, relativement à chacun de ces deux vecteurs, guide sa classification en bon ou en mauvais résultat pour un utilisateur donné. Dans (Lang, 1995), l'idée est exactement la même si ce n'est que l'auteur se base sur une échelle de jugement des articles à cinq niveaux, allant de très bon à très mauvais. Ce sont donc cinq vecteurs prototypiques, représentant chacune des cinq classes, qui sont construits pour modéliser le profil d'un utilisateur. Dans ces deux approches, la construction d'un vecteur prototypique d'une classe donnée s'obtient en faisant la moyenne des vecteurs des instances de cette classe.

Nous évoquions précédemment que Syskill & Webert, décrit dans (Pazzani et al., 1996), opérait une sélection des termes de l'espace. À partir d'évaluations explicites opérées par les utilisateurs, ils disposent d'un ensemble d'articles jugés bons, et d'un autre d'articles jugés mauvais. En se basant sur une mesure de gain d'information, ils parviennent à extraire les termes représentatifs de la classe des bons résultats. Leur méthode s'inspire largement de la fonction de pondération *TF*IDF*, en attribuant un fort poids aux termes particulièrement présents dans les articles de la classe des bons résultats, et parallèlement assez rares dans ceux de la classe des mauvais résultats.

Toutes les approches présentées ici se contentent d'exploiter le seul contenu des pages parcourues par l'utilisateur, sans ajout de données externes. Plus récemment, avec l'essor des réseaux sociaux et l'activité grandissante des utilisateurs sur ces applications, certains auteurs se sont intéressés à l'exploitation des données de ces sources. Dans les travaux de (Phelan et al., 2009), le système suit le parcourt d'un utilisateur donné sur le réseau *tweeter* dans le but de mieux définir ses préférences. Les auteurs considèrent que la consultation volontaire d'un *tweet* particulier est un indice du fait que l'utilisateur ait été intéressé par son contenu. Les *tweets* parcourus par un utilisateur sont donc récupérés pour enrichir la représentation de ses centres d'intérêt.

Malgré quelques variations au niveau de la pondération des termes ou du nombre de vecteurs construits, c'est de nouveau la structure de vecteurs de traits qui prime dans la littérature pour la représentation des profils utilisateur. Le constat est peu surprenant

puisque les deux structures doivent être comparables afin de pouvoir mesurer la pertinence d'un article relativement à un profil utilisateur. Les travaux évoqués dans ce chapitre démontrent en revanche une certaine diversité dans les approches développées pour opérer cette comparaison.

3.3.3 Filtrage d'information

La recommandation basée sur le contenu est une sous-tâche particulière de la tâche de filtrage d'information : l'objectif est en effet de sélectionner les items d'une collection correspondant à un profil utilisateur, et d'exclure les autres. En ce sens, elle est régulièrement reconsidérée comme une tâche de classification dont le nombre de classes varient selon l'échelle de jugement proposée aux utilisateurs.

L'approche la plus basique consiste en une classification binaire, distinguant strictement les bons items des mauvais, mais nous évoquions précédemment des travaux intégrant des échelle de jugements à plus de deux niveaux. (Lang, 1995) implémente dans son système NewsWeeder une comparaison à base de calcul de cosinus. Le vecteur d'un article est comparé à chacun des cinq vecteurs prototypiques de classe en mesurant le cosinus de l'angle qu'il forme respectivement avec chacun d'eux. Il est finalement associé à la classe pour laquelle il obtient le plus haut score de cosinus, et se voit assigner le niveau de pertinence correspondant à cette classe.

Les auteurs de (Kamba et al., 1995) font également appel à un modèle vectoriel, en s'inspirant des méthodes de RI. Ils considèrent le profil utilisateur comme une requête soumise au système, pour laquelle il doit retourner les documents de la collection qui y répondent au mieux.

L'approche proposée dans (Pazzani et al., 1996) s'appuie quant à elle sur des modèles plus classiques de classification. Les auteurs présentent une série d'expérimentations destinées à comparer les performances de quelques uns des algorithmes les plus répandus pour cette tâche (notamment *Naive Bayes*, les k plus proches voisins, un perceptron et une rétropropagation de gradient). Notons qu'ils distinguent dans leur corpus les contenus en fonction des sujets qu'ils abordent et opèrent leurs expérimentations indépendamment pour chacun de ces sous-corpus. Ils observent que les performances des algorithmes sont très variables en fonction des données exploitées pour l'apprentissage, tant au niveau des sujets traités qu'à celui du nombre d'instances utilisées pour cette phase. Une analyse croisée de l'ensemble des résultats obtenus leur permet malgré tout de conclure que *Naive Bayes* est globalement meilleur que tous les autres. C'est donc finalement ce modèle qui est intégré à leur système Syskill & Webert.

Une étude plus récente (Qiu et al., 2009) propose de reconsidérer tout autrement la

tâche de recommandation d'actualités, en s'inspirant de méthodes liées au domaine du *Topic Detection and Tracking* présenté au Chapitre 2. En citant (Sarwar et al., 2001), les auteurs commencent par rappeler qu'un système de filtrage d'information basé sur l'item plutôt que sur l'utilisateur est généralement plus efficace, à la fois en termes de précision des résultats et de complexité algorithmique. Partant de ce constat, ils proposent d'exclure complètement l'utilisateur de leur chaîne de traitement, en construisant un système uniquement basé sur les informations décrivant les articles, cherchant à regrouper les plus proches dans des groupes homogènes.

3.4 Évaluation

Contrairement aux précédentes tâches abordées dans cet état de l'art, la tâche de recommandation ne connaît pas de protocole d'évaluation bien défini. La diversité des items considérés et des objectifs à remplir compliquent la définition d'un cadre strict dans lequel pourraient s'inscrire tous les systèmes développés dans ce domaine.

Malgré cette absence de consensus, certains paramètres sont unanimement considérés comme essentiels pour l'évaluation des systèmes de recommandation. C'est notamment le cas des retours des utilisateurs quant aux items soumis, qui constituent les résultats de référence auxquels comparer les résultats automatiques. On distingue généralement deux types d'indice comme témoins de la satisfaction d'un utilisateur relativement à un item donné : les indices explicites et les indices implicites.

Les premiers nécessitent un investissement de la part de l'utilisateur, à qui l'on demande d'attribuer une note à l'item proposé. Comme nous l'avons constaté dans ce chapitre, l'échelle de notation peut être binaire (Billsus et al., 1999) ou plus étendue (Lang, 1995). D'autres systèmes proposent par ailleurs aux utilisateurs de rédiger un commentaire au sujet de l'item proposé, afin de saisir plus finement ses critères de satisfaction (Resnick et al., 1994). Ce type de jugement explicite est idéal pour représenter les préférences des utilisateurs, mais ceux-ci ne sont pas toujours enclin à fournir cet effort d'annotation. La seconde catégorie des retours implicites intervient alors comme une alternative moins coûteuse, se basant sur les actions des utilisateurs sur une page donnée. En exploitant des critères tels que le temps passé sur la page, sa vitesse de défilement ou encore son enregistrement dans les *Marques-Pages* du navigateur, le système est capable de déduire l'intérêt que l'utilisateur lui a porté. L'absence d'intervention humaine offre un avantage certain à ce type d'indices, bien qu'il soit nécessaire d'en nuancer la fiabilité. Plusieurs facteurs peuvent intervenir comme biais dans la récupération de ces informations. (Lops et al., 2011) présente par exemple une situation dans laquelle l'utilisateur lit un contenu sur son téléphone, sur lequel il reçoit un appel auquel il répond. L'indice de temps de

lecture n'est donc pas exploitable dans ce cas.

Un autre biais de ce type d'évaluation est qu'il est difficile de savoir sur quelle(s) caractéristique(s) de l'item le jugement de valeur est porté. Les auteurs de (Adomavicius, Manouselis et al., 2015) définissent en effet la recommandation comme une tâche multi-critères pour laquelle les préférences sont difficiles à cerner. Considérons par exemple un système de recommandation de musique, établissant ses recommandations sur les critères d'artiste, de genre, de label, de décennies de sortie et de rythme. Imaginons un utilisateur auquel le système propose, sur la base de son historique de consommation musicale, un item particulier. Que l'utilisateur juge binairement cet item comme acceptable ou non, comment savoir sur quelle(s) caractéristique(s) de l'item il a basé son jugement ? L'idéal serait qu'il fournisse lui-même le détail de ses choix, mais l'investissement qu'une telle démarche demande fait d'elle une utopie dans un contexte aussi dynamique que le Web. Ce n'est qu'en croisant les jugements attribués à différents items de la collection que le système peut déduire des tendances quant aux préférences d'un utilisateur.

Par ailleurs, en plus de cette complexité liée aux multiples dimensions des items, une seconde intervient au niveau de la définition des objectifs de l'application de recommandation développée. Dans (Gunawardana et al., 2015), les auteurs soulignent qu'un système de recommandation renferme différentes propriétés, pas nécessairement complémentaires, qui affectent l'expérience de l'utilisateur. Parmi elles, on peut citer l'exactitude, la couverture, la sérendipité et la diversité des résultats, ou encore la modularité, la robustesse et le temps de calcul du système. Toutes ne pouvant être optimisées simultanément, ce n'est qu'en décidant en amont quelle(s) caractéristique(s) de l'application on souhaite évaluer que l'on peut envisager quel protocole d'expérimentation mettre en place.

Dans un cadre industriel, la prise en compte de l'activité du système en conditions réelles d'application est un paramètre important de l'évaluation. C'est encore davantage le cas lorsqu'il s'agit de traiter des données du Web, à la fois riches et dynamiques, dont les caractéristiques influencent fortement les performances du système. Dans un tel contexte, l'idéal est de tester le système directement en ligne, en observant l'influence de ses recommandations sur ses utilisateurs réels. Des méthodes d'A/B testing sont régulièrement utilisées dans de telles configurations, dans l'idée de pouvoir comparer les performances de différents systèmes sur des données similaires (Kohavi et al., 2009 ; Amatriain, 2013).

Le déploiement d'un tel processus d'évaluation étant lourd et coûteux, certains travaux lui préfèrent un processus d'évaluation hors-ligne, à base d'ensemble de test. La construction d'un tel ensemble nécessite de figer la collection d'items à un instant donné, ainsi que de solliciter un échantillon d'utilisateur pour l'évaluation des items sélectionnés. L'ensemble finalement récupéré sert de données de référence auxquels les différents systèmes implémentés peuvent se comparer pour témoigner de leurs performances. L'un des biais majeurs de cette méthode est qu'elle simule l'interaction entre l'utilisateur et le

système, qui constitue pourtant un facteur important dans la tâche de recommandation. Dans une tâche de recommandation, l'objectif est en effet plus de satisfaire les utilisateurs que d'atteindre une haute précision *système* (Konstan et al., 2012). Dans (Clough et al., 2013), les auteurs proposent un ensemble de test dédié à la tâche de recommandation d'actualités, en soulignant qu'il est primordial dans ce type de protocole de ne pas négliger l'utilisateur ni son rôle dans la collecte d'informations de référence.

Des méthodes intermédiaires aux précédentes permettent de pallier cette contrainte en impliquant un ensemble d'utilisateurs dans l'évaluation du système en conditions réelles : il s'agit des tests utilisateur. L'idée est ici de recruter un groupe d'utilisateurs auxquels on demande d'effectuer un certains nombre de tâches en interaction avec le système, comme de juger ses recommandations ou de répondre à certaines questions qualitatives. Afin d'obtenir des données suffisamment représentatives pour être exploitables, un investissement conséquent est nécessaire de la part des utilisateurs, impliquant un certain coût dans la mise en place d'une telle procédure.

En 2013, un atelier de recherche autour de ces problématiques d'évaluation est organisé à l'occasion d'une série de conférences ACM, *News Recommender System and Challenge*[1] (*ACM RecSys'13*). À cette occasion, les systèmes proposés par les différents participants sont évalués en conditions réelles d'utilisation, en interagissant avec des utilisateurs réels dont on exploite les retours (*user feedback*) pour mesurer les performances. En mesurant, sur plusieurs semaines précédant la conférence, le taux de clic par requête de recommandation pour chacun des systèmes participants, il est possible d'évaluer et de comparer leurs performances respectives (Tavakolifard et al., 2013).

En amont de l'évaluation, les participants ont accès à un très large corpus de données sur lesquelles ils peuvent baser leurs développements et expérimentations : le corpus *plista* (Kille et al., 2013). Ce corpus est le fruit d'une collaboration entre l'entreprise allemande plista GmbH[2] et l'université technique de Berlin[3]. De ce fait, l'ensemble des articles composant ces données sont en allemand. À l'instar de Mediabong, plista propose un service de recommandation de contenus et de publicités pour plusieurs milliers de sites web (Brodt, 2013). Afin de promouvoir leur activité et la recherche dans le domaine de la recommandation de contenus, ils proposent à la communauté scientifique un corpus regroupant plusieurs millions d'interactions entre utilisateurs et articles d'actualité issus de différents sites et portails.

Ce principe de *living lab*, dans lequel des utilisateurs réels sont placés au cœur du processus de recherche, a depuis été repris par le programme CLEF (*Conference and Labs of the Evaluation Forum*) qui propose depuis 2014 une campagne dédiée à la tâche

1. https://recsys.acm.org/recsys13/nrs/
2. https://www.plista.com/
3. http://www.tu-berlin.de/menue/home/

de recommandation d'actualités : NEWSREEL[4] (*News Recommendation Evaluation Lab*) (BRODT et HOPFGARTNER, 2014 ; HOPFGARTNER, KILLE et al., 2014 ; HOPFGARTNER, BRODT et al., 2015). Les performances sont alors mesurées en termes d'impressions, de clics et de CTR (*click-through rate*) par jour. Une impression est créée lorsqu'un utilisateur lit un article et que le participant reçoit une requête de recommandation pour cet article. Un clic correspond au fait qu'un utilisateur ait suivi un lien de recommandation proposé lorsqu'il lisait un article. Le CTR représente quant à lui le ratio de clics sur le nombre d'impressions.

L'ensemble des métriques proposées lors de ces campagnes évaluent la performance *système* des moteurs de recommandation en compétition, la satisfaction d'un utilisateur étant simplement déduite des jugements de pertinence qu'il fournit. Toutefois, en tant qu'humain, l'utilisateur peut être influencé par de multiples facteurs lors de sa prise de décision quant au jugement à attribuer à un item, et ces facteurs, non pris en compte dans le cadre d'évaluation, peuvent impacter les performances du système. Certains chercheurs proposent donc de placer l'utilisateur au cœur du processus (*user-centric approach*), à la fois pour le développement du système (MCNEE et al., 2006) et pour son évaluation (PU et al., 2011 ; PU et al., 2012). Afin de modéliser le comportement des utilisateurs, il faut d'abord le comprendre et pour se faire, il est nécessaire de mener des évaluations dans un cadre empirique qui considère l'ensemble des facteurs influençant l'expérience utilisateur (KNIJNENBURG, MEESTERS et al., 2009 ; KNIJNENBURG, WILLEMSEN et al., 2012). En plus de considérer les préférences des utilisateurs dans le processus de recommandation, ils proposent ainsi d'intégrer les facteurs influençant leur *expérience utilisateur*, c'est-à-dire l'évaluation subjective qu'ils font de leur interaction avec le système. Ces facteurs peuvent concerner un aspect du système différent de la pertinence comme la diversification des résultats (WILLEMSEN et al., 2011) ou la sérendipité (YU et al., 2017). Ils peuvent aussi relever d'aspects plus personnels, liés à l'expérience de l'utilisateur, tel que son degré d'expertise quant aux items recommandés (KNIJNENBURG, REIJMER et al., 2011) ou encore à son attachement à la protection de sa vie privée (TELTZROW et al., 2004).

Nous ne détaillerons pas plus ici les différentes métriques proposées pour chacune des propriétés précédemment citées. Le lecteur intéressé pourra se référer à (GUNAWARDANA et al., 2015), proposant un large panorama des problématiques évoquées, et des propositions pour leur résolution. Plus que la présentation d'un protocole et des métriques d'évaluation, nous souhaitions ici mettre l'accent sur les difficultés que cette tâche représente et sur la diversité des propositions présentées dans l'état de l'art. Il apparaît que les systèmes de recommandation, très populaires dans le monde industriel comme dans le monde académique, sont difficilement comparables entre eux, du fait qu'ils ne traitent pas les mêmes données, ne visent pas les mêmes objectifs et n'engagent pas les mêmes

4. http://www.clef-newsreel.org/

utilisateurs. Chacun développe ainsi le protocole qui lui permet d'évaluer et de comparer différentes implémentations de son propre système. Dans le cadre de cette thèse, ces questions autour de l'évaluation nous ont beaucoup interrogées, et nous avons élaboré un protocole adapté aux contraintes posées par le besoin complexe de MEDIABONG, présenté au Chapitre 5.

3.5 Conclusion

Cet état de l'art sur les méthodes de recommandation basée sur le contenu a permis de dégager des pistes de recherche répondant à certaines de nos problématiques.

Après une présentation succincte de la tâche et de ses objectifs, nous sommes revenus plus en détails sur la façon dont elle est généralement modélisée dans les travaux s'y consacrant. La recommandation peut être considérée comme une sous-tâche particulière de la recherche et du filtrage d'information, qui s'en distingue par la prise en compte de l'utilisateur dans son formalisme. Nous en avons présenté le modèle élémentaire, impliquant trois paramètres de base que sont un ensemble d'utilisateurs, un ensemble d'items et un espace d'évaluation.

Nous nous sommes ensuite focalisés sur les systèmes particuliers de recommandation d'actualités qui implémentent pour la plupart des solutions hybrides mêlant approches à base de contenu et approches à base de filtrage collaboratif. Nous intéressant, dans le cadre de cette thèse, uniquement aux méthodes à base de contenu, nous avons détaillé les trois modules impliqués dans ce type de système. Dans un telle application, les articles et les utilisateurs sont tous deux représentés sous la forme de vecteurs de termes pondérés. Nous avons présenté les différentes stratégies proposées pour mesurer la similarité entre ces vecteurs, relevant de méthodes de RI ou d'algorithmes de classification.

Le chapitre se clôt par une discussion sur les méthodes d'évaluation de cette tâche de recommandation. Plutôt que d'en détailler la multitude d'approches et de métriques disponibles dans l'état de l'art, nous en avons exposé les principales difficultés, notamment dues à la multidimesionnalité des données exploitées et des objectifs à remplir.

Mis en perspective avec nos propres problématiques quant à l'évaluation des performances du système, cet état de l'art nous a guidé dans l'élaboration d'un protocole personnalisé. Face à un besoin complexe, un système automatique se voit obligé d'opérer certains choix dans l'optimisation des propriétés qui le définissent, rarement compatibles entre elles. Ainsi, si l'une est privilégiée lors de l'implémentation, les résultats sur les autres propriétés risquent d'être impactés négativement.

Ce qui ressort finalement de ce tour d'horizon est qu'il est nécessaire de trouver un

compromis entre toutes les exigences initialement formulées, afin de proposer un système capable d'y répondre au mieux. C'est ce que nous nous sommes efforcés de faire lors de nos réflexions sur la façon d'évaluer les performances du système développé dans ces travaux, puis leurs applications.

Troisième partie

Contributions de la thèse

Chapitre 4

Semiabong, un système d'appariement de contenus textuels médiatiques

---

La problématique de l'appariement d'articles en ligne et de vidéos a été reconsidérée dans ces travaux comme une tâche de recherche d'information, où l'article correspond à la requête soumise au système, et la vidéo appariée aux résultats qui se limitent à un unique document.

La littérature du domaine a largement fait apparaître qu'un tel système doit gérer deux principales difficultés que sont la synonymie et la polysémie des termes soumis en requête. Le fait qu'un même terme puisse référer à différentes entités, et qu'inversement une même entité puisse être désignée par différents termes, nécessite de considérer un niveau de représentation plus linguistique que brut. Dans ce travail, ces problématiques dépassent le cadre du mot puisque les requêtes traitées sont des textes complets : il nous faut donc gérer ces ambiguïtés aux niveaux supérieurs de la phrase et du texte. Un même sujet peut en effet être traité de différentes manières, depuis différents points de vue – *cf.* **(1)** – et à l'inverse différents sujets peuvent être décrits par des contenus similaires – *cf.* **(2)**. Nous nous sommes attachés à modéliser au mieux les contenus afin de dépasser ces ambiguïtés, tout en répondant aux contraintes industrielles de construire un système efficace nécessitant peu de ressources, essentiellement basé sur les données elles-mêmes.

(1)
Article (2016-03-02) - *Coupe de France : le PSG ira a Lorient en demi-finales*

Video (2016-03-02) - *Coupe de France : le PSG domine Saint-Etienne en 1/4 de finale*

(2)
Article (2016-04-05) - *Metz : condamné pour s'être filmé en prison avec l'application Periscope*
Video (2015-01-31) - *Marseille : un détenu des Baumettes se filme en direct*

Afin de répondre au mieux au besoin initial de MEDIABONG qui souhaitait automatiser cette tâche d'appariement dans le but de minimiser l'intervention humaine, nous avons développé un système, SEMIABONG, dont les étapes sont détaillées dans ce chapitre. Nous présentons tout d'abord les données particulières traitées dans ces travaux (Section 4.1) avant d'exposer les différentes étapes de l'implémentation du système. La Figure 4.1 en présente l'architecture générale, où l'on peut distinguer deux modules :

— Le SRI prend en entrée un article et propose en sortie un ensemble de vidéos qu'il estime proches de cet article. Ce module suit une chaîne classique d'indexation documentaire (Section 4.2), de sélection de candidats, puis de calcul de différents scores pour chacun des candidats retenus (Sections 4.3 & 4.4).

— L'ordonnancement des résultats prend en entrée le précédent ensemble de vidéos et en propose un tri en fonction des scores obtenus par chacune d'elles. Ce module est construit sur la base d'un ensemble de jugements utilisateurs (Section 4.5) utilisés comme exemples pour l'apprentissage d'un classifieur modélisant les besoins spécifiques des utilisateurs du système (Section 4.6).

Nous concluons ce chapitre en Section 4.7, avant de présenter au chapitre suivant la méthode d'évaluation mise en place pour l'évaluation de ce système.

4.1 Données traitées

Nous présentons ici les spécificités des documents traités dans ces travaux à savoir les vidéos et les articles de presse. Bien que MEDIABONG soit désormais ouvert à l'International, ces travaux de thèse se sont exclusivement consacrés aux partenariats nationaux, toutes les données traitées ici sont donc en français.

4.1.1 Vidéos

Les vidéos reçues des producteurs partenaires de MEDIABONG sont accompagnées de descriptions textuelles structurées en un titre, un résumé et éventuellement une liste de mots-clés plus ou moins correctement renseignés selon les producteurs, voire pas renseignés du tout. Il ne s'agit donc pas d'analyser le flux de parole énoncé dans la vidéo, notre analyse se focalisant strictement sur les données textuelles à disposition. Ces vidéos sont, dans notre contexte de recherche d'information, les documents qui forment la collection interrogée par le système et donc les résultats potentiels associés aux articles.

De par l'étendue des partenariats producteurs, qui évolue sans cesse au gré des arrivées et départs de chacun, la collection de vidéos est riche et variée. Nous avons ainsi essentiellement affaire à des contenus d'actualité politique, économique, sociétale et spor-

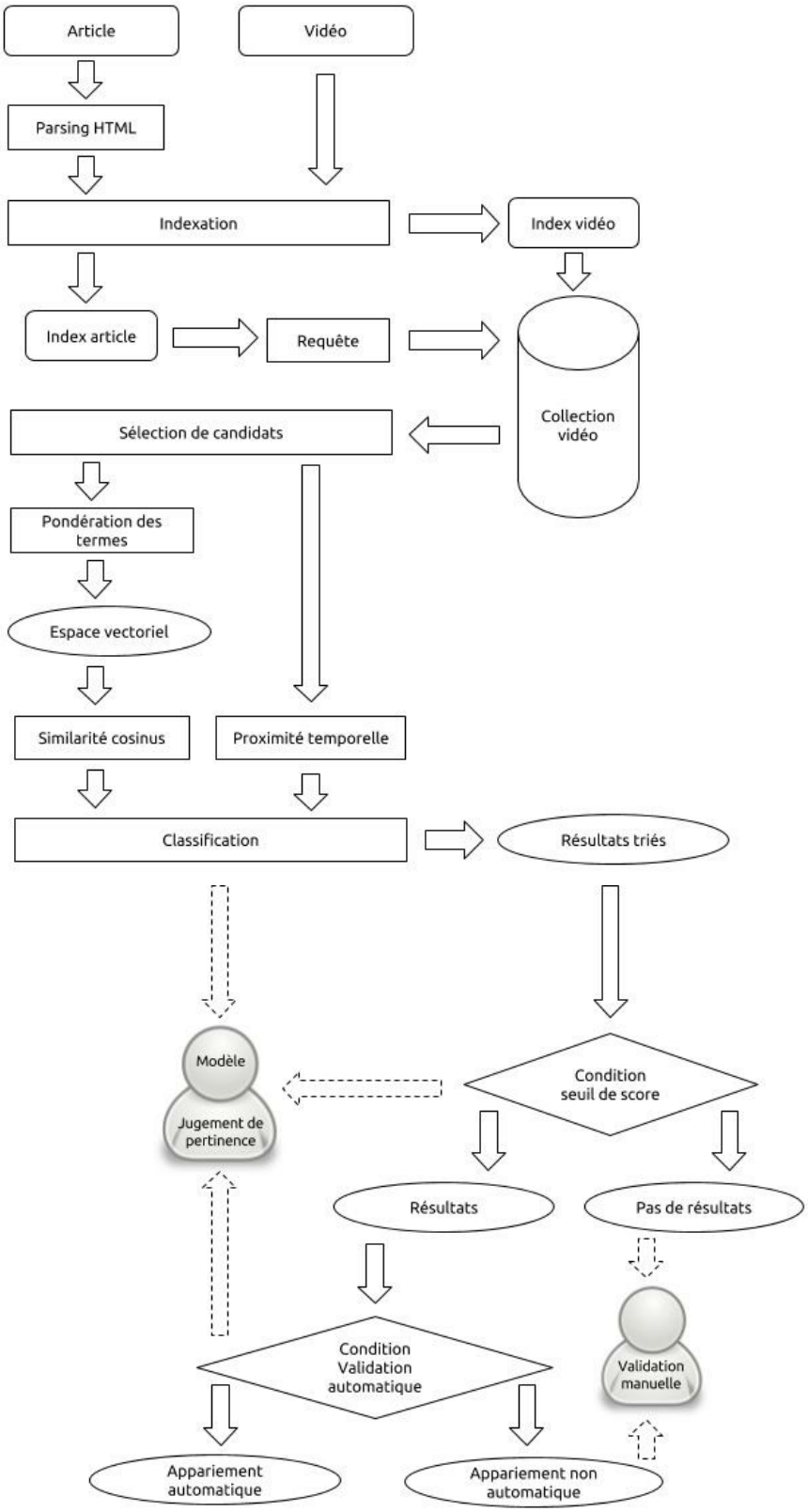

FIGURE 4.1 – Architecture du système SEMIABONG

tive d'étendue nationale et internationale, avec des partenaires tels que *Euronews, TF1, BFM, AFP, France 24* ou *Europe 1*. Parallèlement, mais dans une moindre mesure, nous recevons des vidéos de producteurs plus spécialisés, notamment en presse *People* (*Fashion TV, Zoomin, Public*) ; en santé (*Doctissimo*) ; en cuisine (*750g, Ptitchef*) ou encore en nouvelles technologies (*Journal du Geek*). En moyenne, ce sont 452 nouvelles vidéos qui sont quotidiennement ajoutées à la base de données existante, qui comptait en début de projet un peu plus de 165 000 entrées. À la date du 30 mai 2017, ce sont plus de 635 000 vidéos qui peuplent cette collection en constante évolution. L'Annexe B en présente un échantillon, représentatif de la variété des vidéos traités.

En moyenne, les titres des vidéos comptent 10 termes [1], pour un écart-type de 3, ne révélant que de faibles écarts à cette moyenne. Quant aux résumés, leurs longueurs est plus que variable en fonction des producteurs, en témoigne un écart-type de 67 pour une longueur moyenne de 52 termes. En observant les deux extrêmes, on note en effet que le plus court des résumés fournis ne compte qu'un seul terme, quand le plus long en compte plus de 3 000. Si cette faible longueur, qui touche plus d'une vidéo, révèle un mauvais renseignement du champ de la part du producteur associé, on a pu observer que les résumés les plus longs étaient générés via des outils de *speech-to-text*, transcrivant le contenu oral de la vidéo en texte écrit. *Euronews* est de ceux-ci, et peut produire des vidéos d'une vingtaine de minutes, sachant qu'en une seconde, quasiment 3 termes sont transcrits. On arrive donc facilement à plus de 3 000 mots en sortie.

En plus de ces données textuelles, chacune des vidéos se voit associée des métadonnées concernant sa date de création et le producteur à son origine. Par ailleurs, leurs sont manuellement attribués un ou plusieurs thèmes issus d'une nomenclature propre à MEDIABONG [2], par une de ses équipes en charge de la gestion des producteurs et contenus vidéos.

4.1.2 Articles

Les articles correspondent aux donnés d'entrée du système développé. Ayant reconsidéré celui-ci comme un SRI, ces articles constituent les requêtes soumises, pour lesquelles il s'agit de retrouver les vidéos de la collection y répondant. Ces articles nous arrivent sous la forme d'URLs, desquelles sont extraites les sections de textes souhaitées, *i.e.* données textuelles telles que le titre, le corps et éventuellement un chapô (lorsque la structure de la page intègre un tel champ) et métadonnées telles que la date de publication et le nom du site diffuseur. Cette extraction se fait au moyen d'un *parser* HTML existant, à

1. Le décompte des termes se fait sur une simple segmentation du champ sur l'espace. Ces fréquences de termes sont donc indicatives, mais à nuancer du fait que nous appliquons aux contenus des traitements plus poussés lors de l'indexation.
2. *cf.* Annexe C pour le détail de cette nomenclature.

savoir *Goose*[3], qui s'adapte aux différentes structures de sites pour récupérer le contenu des balises initialement spécifiées. L'Annexe A recense une série de dix exemples d'articles traités par SEMIABONG, où sont présentés les contenus extraits par *Goose*.

En fonction de l'activité des diffuseurs, il peut être donné à traiter au système entre quelques centaines et quelques milliers d'articles par jour. Ce chiffre évolue au rythme de nouveaux partenariats diffuseurs qui s'établissent ponctuellement, ainsi que, parallèlement, à l'arrêt de certains autres. La moyenne se situe aux alentours de 800 articles par jour depuis le début du projet en fin d'année 2014. Ce chiffre reste toutefois à nuancer (l'écart-type à cette moyenne en témoigne clairement, puisqu'il atteint la valeur de 821) du fait que nous travaillions à ce moment-là avec des partenaires tels que *Portail Free* ou *Planet.fr*. Ces agrégateurs puisent leurs contenus de différentes sources sur le Web et le nombre de pages reçues quotidiennement de ces diffuseurs pouvaient atteindre plusieurs centaines. Le nombre total d'articles traités chaque jour dépassait alors largement le millier mais ces deux partenaires se sont retirés au début 2016. Depuis, la moyenne tourne plus autour de 200 articles par jour, avec un écart-type de seulement 85 révélant une faible dispersion.

Par ailleurs, les thématiques abordées par tous ces diffuseurs varient, certains proposant de l'information hétérogène, quand d'autres sont spécialisés sur des domaines particuliers. Parmi les plus actifs et les plus connus, on peut citer notamment *20 Minutes, Europe 1, Metro, Atlantico* ou *France Soir*, aux sujets d'actualité nationale et internationale de tous thèmes ; *La République du Centre, La Provence, Midi Libre* ou *Le Progrès* sur des sujets plus locaux ; *Psychologies, ConsoGlobe* ou *RMC Sport* sur des contenus plus spécialisés, respectivement ici en bien-être, consommation alternative et sport.

En moyenne, les titres des articles comptent 11 termes, avec un écart-type de 4. Quant aux corps d'articles, ils comptent en moyenne 316 termes, avec un écart-type de 226. Cette grande variation tient à l'hétérogénéité des partenaires, dont les sites ne sont pas toujours structurés de façon à ce que le *parser* puisse correctement en extraire le contenu. Le plus court contenu de ce champ ne compte en effet qu'un seul terme (révélant une erreur d'extraction) quand le plus grand en compte plus de 5 000.

De la même façon que pour les vidéos, chacun des articles se voit associer une ou plusieurs thématiques de la nomenclature MEDIABONG, constituant une métadonnée supplémentaire dans la représentation de ces documents.

3. https://pypi.python.org/pypi/goose-extractor/

4.2 Indexation des contenus

Afin de représenter les contenus textuels de façon structurée, dans l'objectif d'un traitement automatique, nous proposons un module d'indexation à base d'extraction. Celui-ci se décompose en deux opérations, une première composée de traitements linguistiques permettant la détection des termes simples, une seconde fondée sur une comparaison à une ressource externe propriétaire pour la reconnaissance d'expressions multi-mots (EMM).

Ce module d'indexation est le même pour tous les types de contenus traités dans ces travaux, *i.e.* à la fois les articles et les vidéos. Les termes extraits sont intégrés aux documents formant les entrées des deux collections MongoDB [4] correspondant à ces deux types de documents. Chaque terme est parallèlement intégré à une troisième collection recensant l'ensemble des termes du corpus, qui constitue notre fichier inverse. La Figure 4.2 décrit le schéma global de ce processus d'indexation, dont nous détaillons les étapes à la suite de cette section.

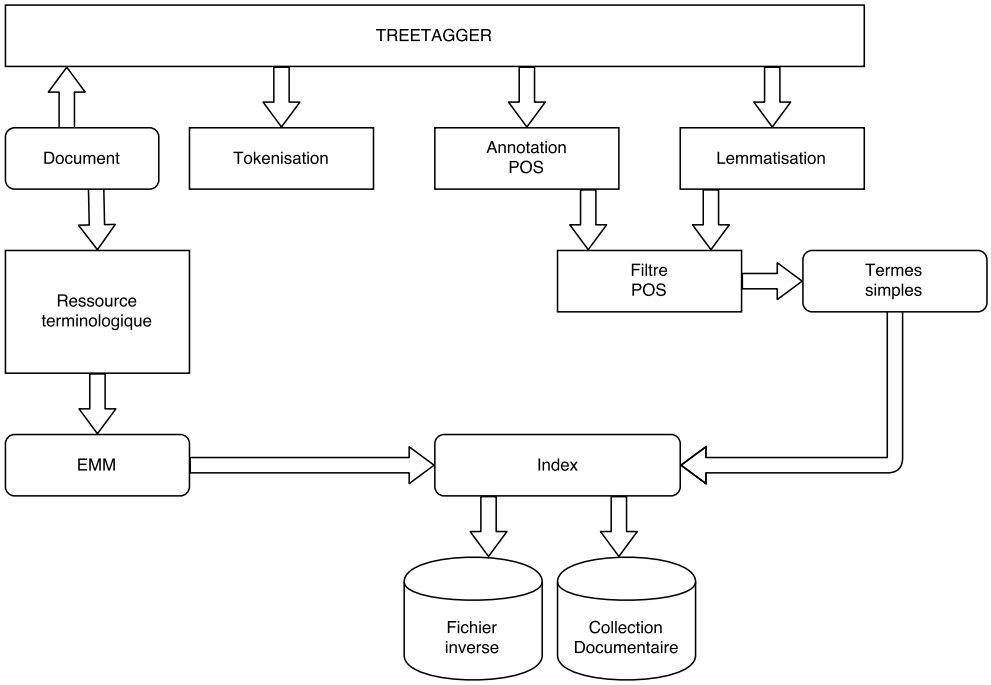

FIGURE 4.2 – Schéma d'indexation des contenus dans SEMIABONG

4.2.1 Extraction de termes simples

La phase d'extraction des termes simples à partir des contenus textuels suit une chaîne de traitement linguistique classique, telle que présentée au Chapitre 1.

4. MongoDB est un système de gestion de base de données non relationnel, que nous avons utilisé dans ces travaux de thèse pour stocker et gérer nos données (https://www.mongodb.com/fr).

Les contenus sont soumis à l'outil TreeTagger que l'on configure pour le traitement du français et procède à leur tokenisation, leur lemmatisation et enfin à leur étiquetage en PoS. Les données de cette dernière opération servent de filtre à l'indexation et permettent de ne retenir que les termes que l'on estime pleins, en opposition au mots-vides que l'on ne souhaite pas voir intégrés à l'index des documents. Par ailleurs, dans le but de normaliser les index, ce sont les lemmes que nous conservons, plutôt que les termes initiaux.

En reprenant la liste initiale présentée dans (Amini et al., 2013) à laquelle nous ajoutons les abréviations, majoritairement référentielles dans nos textes, nous avons choisi dans ces travaux de conserver les termes correspondants aux PoS suivants :

— NOM : noms communs
— NAM : noms propres
— ADJ : adjectifs
— VER* : verbes
— ABR : abréviations

4.2.2 Extraction d'expressions multi-mots

Nous procédons parallèlement à cette recherche de termes simples à une reconnaissance d'EMM, notamment les unités polylexicales dont le sens n'est pas ou peu compositionnel et les entités nommées (EN) polylexicales (Constant, 2012). L'identification des EMM est en effet utile à une représentation plus précise des contenus, pour éviter d'associer des documents sur la base de termes communs qui ne sont en fait que des parties d'expressions dont les référents sont distincts (*e.g.* ***Mont*** *Blanc vs.* ***Mont*** *Saint-Michel ; Édouard* ***Philippe*** *vs.* ***Philippe*** *Katerine ;* ***Front*** *populaire vs.* ***Front*** *National*).

La non-compositionnalité est une spécificité de la langue faisant que certains groupes de mots véhiculent ensemble un sens particulier qui ne peut être interprété à partir de la combinaison des sens de chacun de ses constituants (M. Gross, 1982). Bien que cette notion inclut des expressions à tête verbale, telles que *pleuvoir des cordes*, *se fendre la poire* ou encore *coûter les yeux de la tête*, nous nous intéressons uniquement dans ces travaux aux EMM à tête est nominale. Par exemple, l'expression *liste rouge* doit être considérée comme un tout pour en saisir le sens, parce qu'elle ne réfère en aucun cas à un objet de type *liste* de couleur *rouge*. Pour certaines unités polylexicales, le sens peut être plus transparent, mais leur forme reste malgré tout figée, d'où l'intérêt de les identifier comme des unités à part entière (*i.e. assistant parlementaire, ver de terre*). Par ailleurs, dans le corpus d'actualités qu'est le nôtre, nous rencontrons constamment de nouvelles expressions ainsi formées, aux référents univoques, telles que *mariage pour tous*, *France insoumise* ou *transition écologique*, qu'il est nécessaire de considérer dans leur ensemble.

Les EN sont quant à elles des unités particulières faisant généralement référence à

des personnes, des lieux ou des organisations, qui correspondent traditionnellement à des noms propres (EHRMANN, 2008). Tel que nous l'évoquions dans l'état de l'art, au Chapitre 1, ces unités particulières ne sont pas nécessairement polylexicales, puisque des noms propres de personnes (*i.e. Dalida, Napoléon*), de lieux (*i.e. Marseille, Canada*) ou d'organisations (*i.e. Nasa, ONU*) peuvent ne compter qu'une seule unité. Par ailleurs, deux tâches sont généralement distinguées dans un processus d'extraction d'EN, celle de les identifier d'abord, puis celle de les catégoriser, c'est-à-dire de savoir s'il s'agit d'une personne, d'un lieu ou d'une organisation. Toutefois, nous nous intéressons ici uniquement à l'identification des entités nommées polylexicales, dans le but de lever certaines ambiguïtés lors de l'indexation des documents. S'il nous est par exemple donné un document au sujet de *Manuel Valls*, il est essentiel de considérer ces deux graphies conjointement, pour ne pas associer ce document à un autre qui serait au sujet de *Manuel Noriega*. Avec une représentation classique en *sac de mots*, ces deux documents pourraient être rapprochés parce qu'ils ont en commun le terme *Manuel*, alors qu'aucun lien n'existe entre eux sur la seule base du partage de ce terme. Une telle ambiguïté ne se pose pas pour les entités nommées formées d'une seule unité, qui ne correspondront qu'à une seule dimension dans l'espace des termes, c'est pourquoi nous ne cherchons pas ici à les identifier en tant qu'entités nommées.

Au commencement du projet, MEDIABONG disposait d'une ressource terminologique propriétaire, intégralement construite et mise à jour manuellement. Cette ressource recensait des termes, simples et polylexicaux, jugés par les utilisateurs comme dignes d'un intérêt particulier [5] qu'une intégration à cette ressource traduisait. Dans cette ressource, chaque terme est associé à un type de données, qui peut être PERSONNE, LIEU ou DIVERS, ainsi qu'à un ensemble de thèmes et sous-thèmes de la nomenclature MEDIABONG [6] qui lui sont associés sémantiquement.

Nous sommes donc partis de cette ressource existante, malgré les imperfections qu'elle peut présenter, pour envisager notre système d'extraction d'EMM. Nous avons notamment conscience que d'intégrer à une même ressource des noms de personnes et de lieux n'est pas des plus recommandés, et que de réunir sous une catégorie DIVERS tout ce qui n'est ni personne ni lieu est plus que réducteur. Néanmoins, pour simplifier le développement technique du système global, qui n'aura qu'à interroger cette unique ressource, nous avons fait le choix de conserver la structure telle qu'initialement fournie.

5. *L'intérêt* en question n'est ici ni quantifiable ni modélisable, il s'agissait essentiellement pour les utilisateurs de recenser les termes particulièrement cités à un moment donné dans l'actualité, et à les intégrer à la ressource sur ce seul fait.

6. cf. Annexe C pour le détail de cette nomenclature

Construction de la ressource

Dans le but de réduire l'investissement humain dans l'actualisation dynamique de cette ressource, nous avons développé un module de reconnaissance automatique d'EMM. L'intégration de nouveaux termes dans la ressource est en fait semi-automatique, car elle nécessite encore à ce jour une vérification manuelle des termes après extraction, ainsi que l'attribution du type et du thème précédemment évoqués, que l'outil n'est pour l'instant pas capable de fournir.

Notre approche est partie du constat que les médias ont tendance à tous relater les mêmes faits au même moment, rendant les données récupérées à un instant donné fortement homogènes. C'est en effet ainsi que se crée l'actualité, lorsqu'un média s'empare d'un sujet, que d'autres reprennent ensuite, *etc.*. Nous assistons quotidiennement à ce phénomène dans la presse écrite, et constatons que beaucoup des contenus sont comparables, si ce n'est identiques lorsqu'ils reprennent textuellement des dépêches AFP reçues en continu. Notre méthode d'extraction d'EMM se base sur une approche hybride, mêlant statistiques textuelles et patrons syntaxiques, dont les composants fondamentaux sont les co-occurrences et les segments répétés en corpus tels que décrits dans (Salem, 1986).

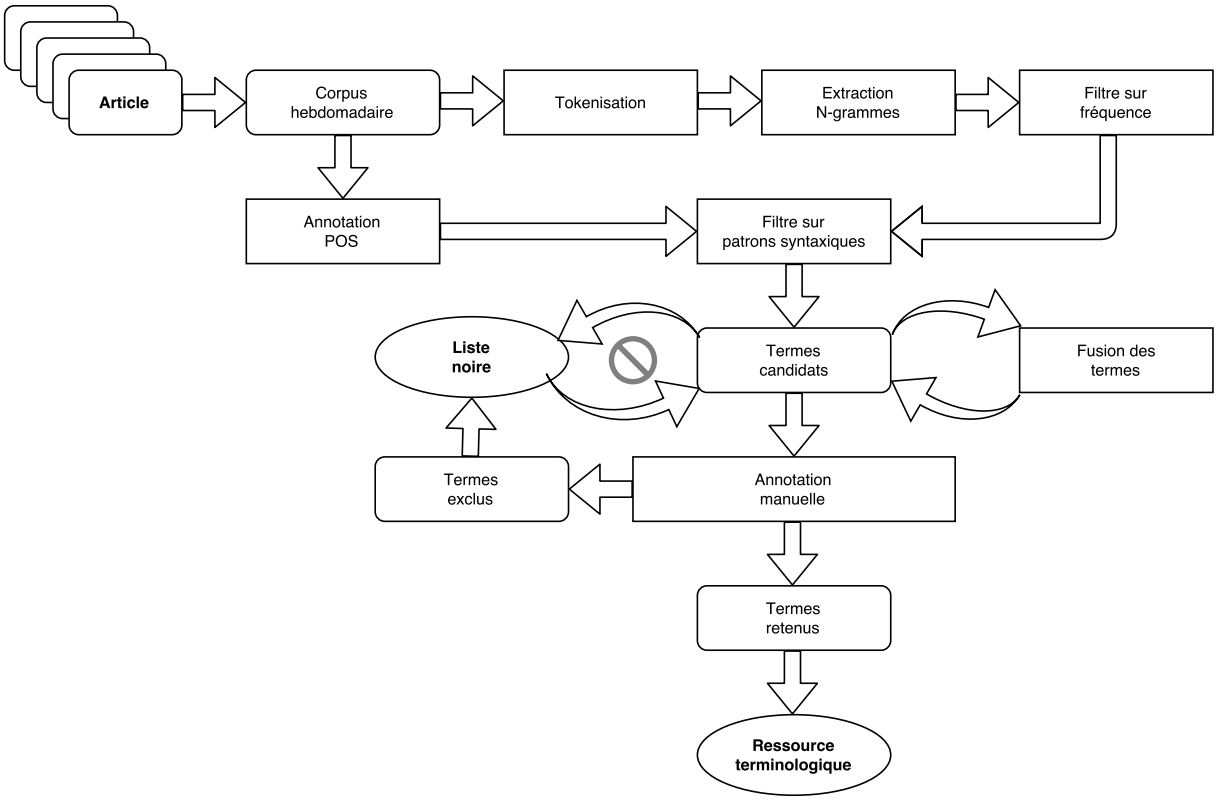

Figure 4.3 – Schéma du processus de mise à jour semi-automatique de la ressource terminologique de Semiabong

Dans un premier temps, nous récupérons l'ensemble des articles publiés sur un intervalle de temps donné. Nous choisissons une fréquence hebdomadaire, qui permet de

récupérer un corpus plus large qu'en travaillant au niveau du jour. Ainsi, chaque lundi de chaque semaine, sont extraits et réunis en un corpus documentaire tous les articles publiés et traités par Semiabong la semaine précédente. Ce corpus est soumis à une chaîne de traitements telle que décrite en Figure 4.3, débutant par une tokenisation et un étiquetage PoS. L'ensemble des bigrammes et trigrammes est ensuite récupéré, *i.e.* tous les suites composées respectivement de deux et trois termes : la Table 4.1 présente un exemple d'une telle extraction à partir d'une chaîne de caractères.

Bigrammes	**Trigrammes**
Zone euro	Zone euro :
euro :	euro : Paris
: Paris	: Paris et
Paris et	Paris et Berlin
et Berlin	et Berlin veulent
Berlin veulent	Berlin veulent aller
veulent aller	veulent aller "
aller "	aller " plus
" plus	" plus vite
plus vite	plus vite plus
vite plus	vite plus loin
plus loin	plus loin "
loin "	loin " dans
" dans	" dans l'
dans l'	dans l' intégration
l' intégration	

Table 4.1 – Bigrammes et trigrammes extraits de la chaîne de caractères « *Zone euro : Paris et Berlin veulent aller "plus vite plus loin" dans l'intégration* ».

Ces *n-grammes* sont ensuite triés sur la base de leur fréquence d'occurrence dans le corpus sélectionné et ne sont conservés pour la suite du traitement que ceux apparaissant dans au moins dix documents différents. Ce premier tri permet de ne retenir que les segments bénéficiant d'un certain degré de figement : la répétition stricte d'un même segment à différents endroits du corpus en est en effet un indice solide. Ce sous-ensemble de *n-grammes* est ensuite soumis à un second filtre, linguistique cette fois. Il s'agit de ne conserver que les segments correspondants à certains patrons syntaxiques prédéfinis. En français, on admet communément (G. Gross, 1988 ; Daille, 1994 ; Mathieu-Colas, 1996) que les patrons caractéristiques sont Nom-Préposition-Nom (***e.g.*** *consigne de vote, cuillère à soupe*), Nom-Adjectif (***e.g.*** *commerce équitable, perturbateur endocrinien*) et dans une moindre mesure Adjectif-Nom (***e.g.*** *premier ministre, fausse alerte*). En fonction de la tâche que l'on souhaite accomplir à l'aide de l'extraction de ce type de données, il est également possible de travailler sur des patrons plus complexes par combinaison de ces patrons élémentaires comme Nom-Préposition-Nom-Adjectif (***e.g.*** *soupçon d'emplois fictifs*) ou par récursivité comme Nom-Préposition-Nom-Préposition-Nom (***e.g.*** *service de répression des fraudes*).

Nous choisissons dans ces travaux de retenir ces patrons syntaxiques classiques de la

langue française auxquels nous ajoutons quelques autres déduits de l'observation du corpus. Nous intégrons notamment la catégorie des abréviations, qui sont monnaie courante dans les contenus de presse relatant des sujets dont les acteurs sont dénotés par des sigles (***i.e.*** SNCF, ONU, CGT).

La Table 4.2 présente l'ensemble des patrons finalement retenus pour la tâche d'extraction d'EMM, ainsi que des illustrations pour chacun d'eux. TreeTagger, qui étiquette les données en PoS, se base entre autre sur des règles graphiques telles que la majuscule pour déterminer la catégorie à assigner à un terme. On peut observer que les termes *Festival* et *Cour* sont ainsi considérés comme des noms propres (Nam). Néanmoins, un terme peut débuter par une majuscule mais être considéré comme un nom commun si l'outil le connaît et le considère comme tel, c'est par exemple le cas de *Ouest* dans *Berlin Ouest*.

Bigrammes		**Trigrammes**	
NOM ADJ	*carte bleue*	NOM PRP NOM	*consigne de vote*
ADJ NOM	*premier ministre*	NAM PRP NOM	*Cour de cassation*
NAM ADJ	*Front national*	NOM PRP NAM	*aéroport de Paris*
NOM NUM	*note 7*	NAM PRP NAM	*Festival de Cannes*
NOM ABR	*grève SNCF*		
NOM NOM	*gaz sarin*		
NAM NAM	*Christiane Taubira*		
NAM NOM	*Berlin Ouest*		
NOM NAM	*usine Goodyear*		
NAM NUM	*France 2*		

Table 4.2 – Liste des patrons syntaxiques utilisés dans Semiabong

Notons par ailleurs que TreeTagger peut se tromper dans l'étiquetage, ce qui peut remonter de mauvais candidats-termes. Il arrive également qu'il ne fournisse aucun lemme à un terme dans les cas où il ne reconnaît pas celui-ci, et lui attribue alors un lemme générique *<unknown>*. Ces cas particuliers sont souvent rencontrés dans le cas de noms propres et/ou emprunts (***e.g.*** *Wolrd Press Photo*) auquel l'outil n'a jamais eu affaire auparavant. Nous exploitons cette information pour également récupérer les segments dont les *n-grammes* de lemmes correspondent à *<unknown><unknown>* et *<unknown><unknown> <unknown>* puisqu'il s'agit régulièrement d'EN.

Avant de soumettre les termes candidats aux utilisateurs pour une vérification puis une annotation manuelle, nous procédons à une étape de fusion itérative des termes. Nous avons fait le choix de limiter l'extraction aux segments de trois termes maximum, mais sommes conscients que certaines expressions référentielles peuvent excéder cette taille. L'idée de cette étape est de comparer tous les segments récupérés partageant au moins un terme plein et d'observer leurs distributions respectives dans le corpus documentaire. S'il s'avère qu'ils apparaissent toujours consécutivement ou si l'un est toujours inclus dans l'autre, alors on fusionne ces deux segments pour en former un plus grand. Par exemple, on observe dans la Table 4.2 le segment *note 7*. Le corpus à partir duquel ce segment est extrait contient également le segment *Galaxy note* (Nam Nom), lui aussi extrait.

En comparant leurs distributions, le système observe que pour toutes les occurrences de *Galaxy note*, le second terme *note* correspond toujours au premier terme des occurrences de *note 7*. Ces deux segments sont donc concaténés pour construire le terme-candidat *Galaxy note 7*. C'est un processus itératif au sein duquel les nouveaux termes construits peuvent à nouveau être comparés au reste des candidats : il est alors possible de se retrouver en fin de traitement avec des termes particulièrement longs (***e.g.*** *abus de faiblesse sur Liliane Bettencourt ; accord du syndicat de médecins fmf ; chef d'état-major des armées*).

L'ensemble des candidats retenus est finalement soumis aux utilisateurs en charge de la sélection de ceux méritant d'être ajoutés à la ressource. Notons qu'un terme exclu par un utilisateur est également stocké dans un fichier (*liste noire*) afin de ne pas le présenter une seconde fois s'il est de nouveau extrait. S'ajoute à cette tâche de sélection une tâche d'annotation, puisque tout terme intégré à la ressource doit être associé à un unique type. Ces annotations constituent des métadonnées sur les termes utiles à la représentation des documents et à la RI, que nous exploitons dans SEMIABONG.

L'Annexe E présente un exemple de sortie de cet algorithme d'extraction d'EMM, ainsi que les annotations manuelles associées aux candidats proposés.

Comparaison à la ressource

La ressource terminologique de MEDIABONG comptait, avant la mise en place de son actualisation systématique en Septembre 2016, 9 125 entrées. Au début du mois de juin 2017, 787 nouveaux termes y avaient été ajoutés, portant le nombre d'entrées à 9 912, avec une moyenne de vingt nouveaux termes par semaine. Parmi eux, 448, soit 57%, se sont vu attribuer la catégorie DIVERS ; 281, soit 36%, la catégorie PERSONNE ; 58, soit 7%, la catégorie LIEU.

Il est difficile d'évaluer la qualité intrinsèque d'une telle ressource, principalement parce que nous ne disposons d'aucun corpus annoté auquel les sorties automatiques pourraient être comparées. Par ailleurs, les objets qui la peuplent sont assez hétérogènes et finalement assez peu définis linguistiquement parlant, compliquant l'idée d'un ensemble de données de référence dont la construction nécessite des règles stables et définies.

Il est toutefois possible de mesurer l'apport d'une telle ressource par une évaluation externe (NÉVÉOL et al., 2006 ; NAZARENKO et al., 2009), c'est-à-dire en comparant les résultats obtenus sur deux types d'index, dont l'un intègre ces données tandis que l'autre non. Le Chapitre 6, présentant une série d'expérimentations menées dans le cadre de cette thèse, propose une telle comparaison.

L'indexation des contenus dans SEMIABONG interroge cette ressource pour récupérer les EMM y figurant. Il s'agit d'une simple comparaison de chaînes de caractères, à base

d'expressions régulières, parcourant la liste des termes de la ressource pour les repérer dans les textes. Lorsqu'un terme correspond à une sous-chaîne de caractères du texte, il est extrait et injecté dans l'index du document correspondant.

Exemple

Pour conclure cette section, voici un exemple décrivant le contenu initial d'une vidéo de notre collection, ainsi que l'ensemble des termes extraits par le module d'indexation que nous venons de présenter.

Titre	*La communauté gay remercie Taubira*
Résumé	*Au sein de la communauté homosexuelle, Christiane Taubira est devenue une icône. L'ancienne garde des Sceaux, qui a démissionné ce mercredi, s'est notamment révélée en portant le texte sur le mariage pour tous. Bon nombre de militants retiennent également sa disponibilité et son écoute.*
Tags	*Christiane Taubira, Démission de Taubira, mariage pour tous, homosexuel, Gay*
Termes simples extraits	remercier, sein, devenir, icône, ancien, démissionner, mercredi, révéler, porter, texte, bon, nombre, militant, retenir, disponibilité, écoute
EMM extraites	communauté gay, communauté homosexuelle, christiane taubira, garde des sceaux, mariage pour tous

TABLE 4.3 – Exemple d'indexation de contenu dans SEMIABONG

L'index final des documents est plus structuré qu'une simple liste de termes présents dans son texte puisque l'on en distingue les sections et les types de termes. Sa structure se décompose en quatre différents ensembles de termes :

— Termes simples du titre
— EMM du titre
— Termes simples du résumé
— EMM du résumé

Ce découpage nous permet dans la suite du traitement d'accorder différents degrés d'importance aux termes en fonction de leurs métadonnées.

4.3 Recherche de contenus similaires

À l'instar de la majorité des systèmes de *Topic Detection and Tracking* présentés au Chapitre 1, traitant des données de presse écrite, nous avons choisi d'implémenter un modèle vectoriel pour l'association de contenus. Cette section en détaille les étapes, de la sélection initiale des résultats candidats à leur attribution de score.

4.3.1 Sélection de candidats

La première étape du processus de RI est de récupérer, parmi l'ensemble des documents de la collection considérée, un sous-ensemble de candidats en réponse à la requête soumise. Cette sélection se fait généralement sur la base de la comparaison des termes de la requête au fichier inverse, à l'issue de laquelle on ne conserve que les contenus partageant au moins un terme de la requête.

Nous avons fait le choix, pour le processus d'indexation, de conserver tous les termes que nous estimions pleins, en ne retenant que ceux dont la catégorie syntaxique appartenait à une liste pré-établie. Malgré cette sélection, certains termes d'index restent peu représentatifs du sens véhiculé par les contenus qui les intègrent, c'est notamment le cas des adverbes *être* et *avoir* qui apparaissent dans une majorité de documents mais qui ne renseignent pas sur le sujet qu'ils développent respectivement. Il est possible d'opérer au niveau de cette sélection de candidats un second filtre d'exclusion de terme, qui peut s'accomplir soit à l'aide d'une *liste noire* des mots-pleins à exclure, soit sur la base de leur fréquence documentaire. La loi de Zipf (ZIPF, 1935) est une des lois fondamentales en RI : elle spécifie que la fréquence d'occurrence d'un terme en corpus est inversement proportionnelle à son rang. Autrement dit, le terme le plus fréquent, *i.e.* celui de rang 1, a une fréquence d'occurrence deux fois supérieure à celui de rang 2 ; 3 fois supérieure à celui de rang 3, *etc.* Sur la base de cette loi, il est possible de restreindre la taille de l'espace des termes considérés, en excluant ceux à la fois trop présents en corpus, révélant une certaine généricité, et les termes très rares, révélant à l'inverse une forte spécificité.

Dans notre contexte, le fichier inverse comptait 269 972 entrées au début du mois de juin 2016. En filtrant les termes aux fréquences extrêmes [7], la taille de cet ensemble diminue plus que de moitié et ne compte plus alors que 106 592 termes. Lors de la sélection des candidats pour une requête, nous appliquons ce filtre sur les fréquences dans le double but de restreindre la sélection aux documents partageant des termes pertinents et de réduire par ailleurs le temps de calcul des traitements suivants, qui dépend de la taille de l'espace de termes considéré. Malgré un tel tri, les ensembles de candidats récupérés pour une requête comptent en moyenne environ 89 000 documents, ce qui reste conséquent pour le système. Or dans notre contexte industriel, la rapidité du temps d'exécution est un facteur presque aussi important que la pertinence des résultats, un article ne pouvant attendre trop longtemps avant de se voir associer une vidéo. C'est pourquoi nous avons fait le choix de limiter l'ensemble de candidats au 2 000 documents partageant le plus de termes avec la requête, garantissant ainsi un temps d'exécution de l'ordre de la seconde. Si ce choix est discutable, de part sa rigidité, il reste néanmoins justifiable du fait que pour chacune des requêtes de l'ensemble de test présenté en Chapitre 6, au moins une

7. Nous choisissons ici de conserver ceux dont la fréquence documentaire se situe dans l'intervalle [2,10 000]

vidéo jugée pertinente se trouve dans les 2 000 candidats de tête. Notre objectif étant de n'associer qu'une unique vidéo à un article, nous avons estimé cette décision mesurée et cohérente.

Les métadonnées des termes d'index nous permettent à ce stade d'opérer une sélection plus fine qu'en se basant sur la simple incrémentation d'un compteur à chaque terme commun. Ainsi, pour le premier terme d'index de l'article soumis en requête, on récupère l'ensemble des documents le contenant via l'interrogation au fichier inverse. On initialise un compteur pour chacun de ces documents, que l'on incrémente de 1 si le terme est un terme simple de la description ; de 2 s'il s'agit d'un terme du titre ou d'une EMM issue de la ressource externe. Les documents sont stockés dans une liste temporaire, puis la même opération est exécutée pour le second terme de l'index, incrémentant possiblement le compteur de certains documents existants, et ajoutant de nouveaux documents à la liste, *etc.* jusqu'au bout du parcours exhaustif des termes d'index de l'article. Notons que les pondérations sont cumulables, c'est-à-dire qu'une EMM présente dans le titre de l'article permettra au compteur des documents le contenant d'augmenter de 2+2 = 4.

L'ensemble de documents candidats ainsi constitué est finalement trié par ordre décroissant de la valeur du compteur, et seuls les 2 000 premiers documents de cette liste sont retenus pour la suite du traitement.

4.3.2 Espace vectoriel et projection des documents

L'espace vectoriel n'est généré qu'à partir des termes composant les 2 000 candidats sélectionnés. Ainsi, pour un article donné, les termes d'index des 2 000 vidéos retenues à ce stade constituent chacun une dimension de l'espace dans lequel tous les documents sont projetés.

Projeter un vecteur documentaire dans cet espace de termes implique d'attribuer à chaque dimension du vecteur, chaque terme donc, une valeur. Nous faisons le choix d'une pondération classique basée sur la fonction *TF*IDF* (*cf.* Chapitre 1 pour une présentation détaillée) que nous adaptons grâce aux métadonnées des termes à notre disposition. De la même façon que lors de la sélection des candidats, une surpondération est appliquée à certains des termes de l'index, en multipliant simplement leur *TF* par un coefficient c correspondant à la somme de deux coefficients $a_{(t,D)}$ et $b_{(t)}$ dépendant respectivement de la position du terme t dans le document D (TITRE ou DESCRIPTION) et du type du terme t (EN_P, EN_L, EN_D ou *simple*).

Formellement, pour tout terme t d'un document D, son poids dans le vecteur de D s'obtient par l'équation 4.1. Le détail des pondérations appliquées aux termes est présenté

en Annexe G, soumise à confidentialité par Mediabong [8].

$$w(t,D) = cTF_{(t,D)} * IDF_{(t)} \tag{4.1}$$

C'est en s'appuyant sur les indices de saillance des titres, présentés en Partie I, que nous avons fait le choix dans SEMIABONG d'accorder un poids supplémentaire aux termes présents dans le titre. Ils condensent en effet généralement l'information véhiculée par l'ensemble du document qu'ils titrent (RINGOOT, 2014). Ce constat se vérifie pour la grande majorité des partenaires diffuseurs avec lesquels nous travaillons, bien que d'autres dans le paysage médiatique français optent au contraire pour des titres plus métaphoriques voire satiriques dont le sens peut être plus figuré. Par ailleurs, en reprenant les conclusions d'études sur le discours médiatiques également présentées en Partie I, nous avons fait le choix d'accorder un poids supplémentaire aux EN de PERSONNE et de LIEU qui situent l'actualité dans un contexte spécifique qu'il nous importe de souligner dans les représentations documentaires (MOIRAND, 2007 ; STEIMBERG, 2012).

Les poids sont donc ainsi calculés pour chacun des termes de l'article ainsi que pour ceux de chacune des vidéos candidates. Les vecteurs résultant de cette opération sont tous projetés dans le même espace de termes, au sein duquel il devient possible de comparer les documents automatiquement. Notons que l'*IDF* est ici calculé relativement au sous-ensemble de documents candidats et non par rapport à l'ensemble de la collection. Cette remarque n'est pas anodine car un même terme peut se retrouver avec des poids très différents selon la collection considérée. En effet, les vidéos sélectionnées le sont sur la base des termes partagés avec l'article et ont donc potentiellement entre elles un ou plusieurs termes communs. Or l'*IDF* permet d'accorder un poids fort aux termes particulièrement rares dans la collection considérée, donc si un terme de l'article, même important, se retrouve dans beaucoup des documents sélectionnés, il n'aura qu'un faible poids dans les vecteurs documentaires par rapport à d'autres termes aux distributions plus sporadiques. C'est en conscience que nous avons fait ce choix, qui permet justement de distinguer parmi ces vidéos celles présentant des termes génériques (par rapport à ce sous-ensemble documentaire), de celles partageant avec l'article des termes plus spécifiques qui pourraient s'avérer, selon nous, être des réponses plus précises à la requête. La fin de cette section présente un exemple d'association article-vidéo démontrant l'intérêt d'une telle approche.

8. Le lecteur intéressé par ces informations est invité à en faire la demande par mail à l'adresse adele.desoyer@gmail.com.

4.3.3 Calcul de similarité

L'ensemble des vecteurs étant projetés dans un même espace vectoriel, il devient possible de les comparer entre eux, via le calcul du cosinus qu'ils forment deux à deux (Salton, Wong et al., 1975). Dans le cadre de la RI, on cherche à comparer le vecteur requête à chacun des vecteurs documents et plus précisément dans notre contexte, le vecteur article à chacun des vecteurs vidéos.

Le cosinus considère à la fois le produit scalaire des vecteurs requête q et document D_i (*cf.* équation 4.2) et le produit de leurs normes, produisant en sortie une valeur rendant compte du degré de similarité entre les deux. Plus l'angle entre les vecteurs est faible, plus le cosinus est élevé, cette valeur étant proportionnelle au niveau de similarité entre les deux vecteurs.

Chacune des paires de documents formées par l'article et chacune des vidéos candidates se voient donc attribuer un *score_thematique*, tel que décrit en 4.2.

$$score\_thematique\ (q,D_i) = \frac{q.D_i}{\|q\|\|D_i\|} \tag{4.2}$$

4.3.4 Exemple

Nous terminons cette section par un exemple extrait de l'ensemble de test décrit en Chapitre 6, présentant un article et les trois meilleurs résultats retournés par le système. Comme pour les précédents exemples de ce mémoire, nous ne présentons ici que les titres des documents, article et vidéos, qui suffisent à l'humain pour faire un lien inter-documentaire. Les données analysées par le système sont en revanche bien toutes celles décrites précédemment dans cette section.

Article	*Usain Bolt s'est entraîné avec une équipe de DH de la Côte d'Azur*			
Vidéos Retournées				
Rang	**Titre**	**Termes simples communs**	**EMM communes**	**Score thématique**
1	*Usain Bolt bientôt au Borussia Dortmund ?*	athlétisme, borussia, dortmund, football	usain bolt	0.51
2	*Usain Bolt joue au foot avec les amateurs du JS St-Jean Beaulieu*	football	usain bolt	0.47
3	*"I am Bolt" le documentaire sur Usain*	athlétisme	usain bolt	0.18

Table 4.4 – Exemple de résultats retournés par Semiabong pour une requête donnée

L'exemple démontre que le système a su accorder un poids particulier au terme *usain bolt* qui est en effet l'acteur du sujet décrit dans l'article, tout en saisissant dans le même temps qu'il était important de retourner des vidéos au sujet de football, et non d'athlétisme. La majorité des vidéos retenues dans l'ensemble de candidats sont au sujet de diverses performances de l'athlète dans sa discipline, mais il était ici essentiel d'éviter de considérer ces documents-là comme résultats optimaux au regard du contenu de l'article. Le fait de construire l'espace vectoriel à partir des 2 000 candidats seulement permet au système d'accorder un poids plus faible au terme *athlétisme* qu'au terme *football*, ce dernier étant bien moins présent dans les documents du sous-ensemble que le premier.

4.4 Considérations temporelles

L'une des problématiques de cette thèse réside dans la double contrainte du besoin d'information, exigeant à la fois une similarité thématique des contenus et une proximité temporelle. De précédentes études ont démontré qu'il existait différents types de requêtes et que la mesure de similarité thématique seule ne pouvait toutes les satisfaire. La fraîcheur des résultats retournés intervient comme critère important à considérer pour des requêtes à forte sensibilité temporelle (DAKKA et al., 2012) et la majorité des contenus médiatiques que nous traitons dans ces travaux, comme dans ceux de (DONG et al., 2010), sont de ces requêtes sensibles au temps. MEDIABONG a d'ailleurs particulièrement insisté sur cette dimension temporelle au début du projet de thèse et souhaitait favoriser les résultats récents. Ainsi, l'exemple **(3)** ci-dessous présente une paire article-vidéo n'ayant pas satisfait MEDIABONG qui attend en réponse à l'article une vidéo décrivant le même événement d'actualité plutôt qu'un événement similaire datant de l'année précédente[9].

(3)
Article (2016-11-21) - *Strasbourg : Le marché de Noël ne sera pas annulé après l'opération antiterroriste*
Vidéo (2015-12-01) - *À Strasbourg, un marché de Noël sous haute sécurité*

Nous détaillons dans cette section la fonction de *score_date* implémentée dans le système SEMIABONG ainsi que la façon dont les deux scores calculés pour une paire article-vidéo sont considérés dans un score global, exprimant à la fois leur degré de similarité thématique et celui de leur proximité temporelle.

9. Il s'avère dans ce cas qu'il y a eu des contre-temps sécuritaires sur le marché de Noël de Strasbourg à la fois en 2015 et en 2016, suite à des menaces d'attentats.

4.4.1 Fonction de *score_date*

La dimension temporelle peut être considérée comme une variable faisant partie du contexte de la requête soumise au système, qu'il faut saisir, modéliser puis intégrer au calcul du score final.

Lors de la définition du besoin initial, MEDIABONG a particulièrement insisté sur cette composante temporelle dans la sélection des vidéos pour les articles. Le système existant dans l'entreprise avant le début de cette thèse n'intégrait pas cette dimension dans son système de recommandation vidéo, alors que MEDIABONG faisait de la fraîcheur des résultats un argument de vente auprès des diffuseurs. Afin de pallier ce biais, MEDIABONG a donc développé une fonction empirique, saisissant le degré de proximité temporelle entre un article et une vidéo, dans le but de favoriser les vidéos très récentes (*i.e.* celles datant du jour-même de la publication de l'article ou de la veille) sans exclure les plus anciennes.

$$score\_date\ (q,D_i) = \sqrt[4]{\frac{1}{log_{10}(\sqrt{I+2})}} * a \tag{4.3}$$

L'équation 4.3 présente cette fonction de *score_date* où I correspond à la distance en nombre de jours séparant la date de publication de l'article passé en requête q de celle de la production de la vidéo D_i, et a varie en fonction de la valeur de I. Ceci permet d'attribuer un plus fort score aux vidéos très récentes, *i.e.* du jour-même ou de la veille : $a = 1.4$ si $I = 0$; $a = 1.2$ si $I = 1$; $a = 1$ sinon.

Bien qu'empirique, et développée par MEDIABONG pour répondre à un besoin industriel précis, cette fonction s'inspire de celle proposée dans (CHY et al., 2015) – *cf.* équation 4.4 – en intégrant une fonction inverse à laquelle MEDIABONG ajoute un logarithme au dénominateur pour lisser les écarts entre vidéos très anciennes et vidéos très récentes.

$$RecencyScore = \frac{1}{(\sqrt{I+1})} \tag{4.4}$$

La courbe de cette fonction est présentée en Figure 4.4, où l'on peut observer une forte décroissance entre les scores attribués aux vidéos datant du jour-même ou de la veille de la publication d'un article et ceux des vidéos plus anciennes. Le logarithme permet ensuite de stabiliser cette décroissance pour ne pas exclure des résultats des vidéos datant de la semaine, du mois, voire de l'année précédant la date de publication de l'article.

Nous affirmions plus haut que la majorité des articles traités relatent des faits d'actualité, dits contenus *chauds*, pour lesquels la récence du résultat est primordiale. Une minorité d'entre eux traitent à l'inverse de contenus *froids*, atemporels, pour lesquels il est peu pertinent de considérer un paramètre temporel dans le calcul du score de similarité. Nous avions initialement envisagé de construire un modèle de classification binaire

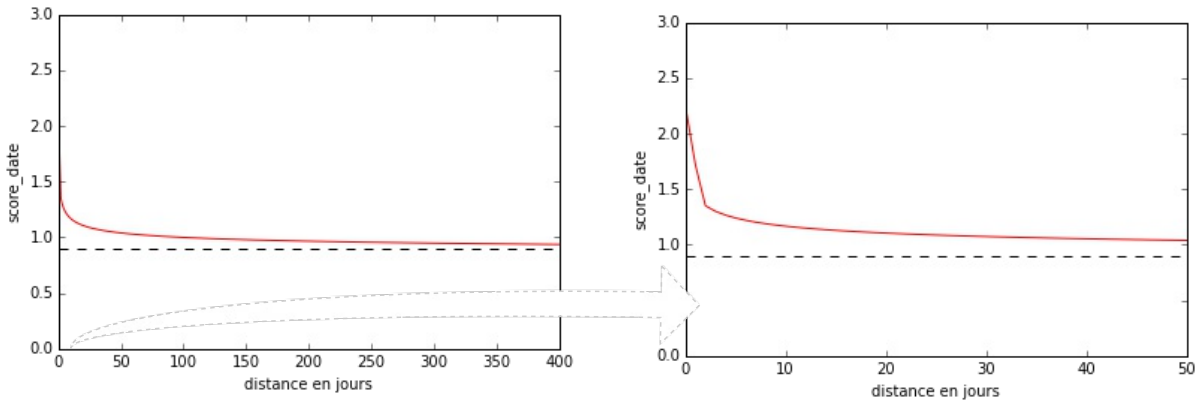

FIGURE 4.4 – Courbe de la fonction *score_date*

permettant de distinguer les articles *chauds* des articles *froids*, dont les sorties nous auraient permis de décider si le paramètre temporel était, ou non, nécessaire au calcul du score de similarité avec les vidéos. Néanmoins, les contenus de chacune de ces deux classes sont fortement hétérogènes et il est rapidement apparu qu'un modèle uniquement basé sur les vecteurs de leurs termes offraient des résultats peu concluants. Nous avons par la suite essayer d'y ajouter des attributs relatifs aux thèmes associés aux contenus, en supposant que les articles *chauds* relatent plutôt des contenus de politique, de sport ou encore d'économie, tandis que les contenus *froids* relèvent plutôt des thèmes concernant la beauté, la santé ou encore la cuisine. Les exemples relevant de cette seconde catégorie étant assez rares parmi l'ensemble des articles traités, l'échantillonnage des données d'apprentissage était fortement déséquilibré lorsqu'on souhaitait disposer d'un ensemble d'exemples suffisamment conséquent pour en déduire des spécificités de classes. En réduisant cet ensemble de données dans le but de rééquilibrer les classes pour l'apprentissage, le modèle s'est avéré peu généralisable, faute d'exemples et de régularités intra-classe permettant de distinguer les deux. Les résultats peu concluants de ces expérimentations en classification nous ont amené à renoncer à une telle distinction de contenus en fonction de leur sensibilité temporelle.

En faisant donc le choix de ne pas distinguer ces deux types de contenus, en n'implémentant qu'une unique fonction de score pour l'ensemble des articles soumis en requête, il était nécessaire de composer une fonction de *score_date* à la fois tolérante vis-à-vis des vidéos (très) anciennes et avantageuse vis-à-vis des (très) récentes. La fonction proposée dans ces travaux de thèse est empirique mais remplit tout-à-fait ces deux contraintes initiales.

4.4.2 Compromis entre les scores

La question qui s'est posée à nous à ce stade de l'implémentation était de savoir comment considérer simultanément les deux scores décrivant la relation entre un article et une vidéo, *score_thematique* et *score_date*, dans un score final rendant compte d'une similarité globale.

En s'inspirant des méthodes de recommandation décrites au Chapitre 2, nous avons mis en place un processus d'évaluation orientée utilisateur, permettant à ceux-ci d'attribuer un jugement de pertinence aux vidéos proposées en réponse à un article. Ces annotations constituent des données utiles pour comprendre leurs besoins particuliers et sont par la suite utilisées comme instances d'apprentissage pour générer un modèle d'ordonnancement des résultats en fonction de leur pertinence vis-à-vis d'un utilisateur particulier.

Néanmoins, contrairement aux systèmes de recommandation qui cherchent à construire pour chacun de leurs utilisateurs un profil spécifique, nous souhaitons à l'inverse construire un modèle générique pour l'ensemble des utilisateurs. Dans notre contexte, l'objectif est en effet d'associer une unique vidéo à un article, censée satisfaire le plus d'internautes possibles. Les données récupérées des sessions d'évaluation ne distinguent donc pas ici les utilisateurs entre eux, ce qui est modélisé est plus de l'ordre d'un profil *macro-utilisateur*, qui associe les préférences de chacun des juges, même s'ils ne sont pas toujours d'accord entre eux (*cf.* Chapitre 6 pour une mesure de l'accord inter-annotateur).

Dans ce cadre, les utilisateurs en charge de l'annotation sont des salariés de MEDIABONG, notamment deux d'entre eux en charge des partenariats avec les producteurs et diffuseurs, particulièrement au fait des contenus gérés par SEMIABONG. Se sont ajoutés à ce binôme, au cours de ces trois années de thèse, d'autres annotateurs au gré des arrivées et départs de stagiaires et alternants dans l'entreprise. Ainsi, s'ils étaient au maximum quatre entre janvier et août 2016, ils n'étaient dans la majeure partie des cas que trois pour cette tâche d'attribution de jugement de pertinence.

Avant de soumettre à ces juges un premier ensemble de résultats, il a fallu décider de la façon d'associer les deux scores calculés. N'ayant *a priori* aucune données décrivant les préférences des utilisateurs, nous avons implémenté le simple produit des deux scores comme score final (*cf.* équation 4.5).

$$score\_final\ (q,D_i) = score\_thematique * score\_date \tag{4.5}$$

Les sections suivantes s'attachent à présenter d'une part la méthode d'évaluation mise en place et d'autre part la façon dont les données récupérées de sessions d'évaluation ont été exploitées pour modéliser les préférences des utilisateurs du système.

4.5 Annotation de corpus

Afin de modéliser les préférences des utilisateurs du système, nous avons mis en place une session d'annotation dans le but de récupérer des données représentant leurs choix quant à l'appréciation d'une vidéo en réponse à un article. Notre objectif est de pouvoir, ensuite, construire un modèle capable de discriminer les bons résultats des mauvais, à partir de ce que l'on connaît des utilisateurs. Il s'agit donc de calculer un classifieur binaire, appris à partir de données de référence, capable de déterminer l'étiquette à attribuer à de nouvelles paires article-vidéo, *i.e.* soit bonne, soit mauvaise. Pour se faire, il est donc nécessaire de construire en amont un corpus annoté manuellement, que nous présentons dans cette section.

Dans la perspective d'une distinction stricte entre bons et mauvais résultats, nous avions initialement envisagé de proposer une annotation binaire de ces données de référence, en proposant aux utilisateurs une échelle de notation à seulement deux niveaux. Toutefois, au regard des différentes exigences du besoin de l'entreprise, il est apparu qu'un utilisateur du système pouvait sélectionner une vidéo sans la considérer comme réellement pertinente, pour garantir le remplissage de la page. Dans ce type de cas particulier, l'utilisateur ne considère la vidéo comme ni vraiment bonne, ni vraiment mauvaise. Or dans le cadre d'une annotation binaire, il est forcé de choisir une de ces deux étiquettes pour chacune des instances à annoter. En ne conservant que ce jeu d'étiquettes, nous prenions alors le risque de nous retrouver avec des instances ne représentant pas réellement la classe à laquelle elles sont associées, et de biaiser l'apprentissage d'un modèle censé représenter chacune des classes à partir des données qui les composent respectivement.

Pour pallier ce risque, nous avons donc préféré proposer aux annotateurs une échelle ternaire, grâce à laquelle il leur était possible d'assigner un jugement qui n'était ni mauvais, ni bon. Ce troisième niveau, intermédiaire aux deux autres, offre une alternative aux juges qui ne sont alors pas forcés de choisir entre les deux notes extrêmes. Grâce à cette méthode, nous espérions garantir la fiabilité des données présentes dans les deux classes extrêmes, qui sont celles que nous souhaitons modéliser ensuite. Ce système d'annotation, dont l'interface est illustrée en Figure 4.5, permet donc aux juges d'attribuer à chaque vidéo proposée en réponse à un article :

— 1 étoile s'ils la considèrent mauvaise (classe 1*)
— 2 étoiles s'ils la considèrent moyenne (classe 2*)
— 3 étoiles s'ils la considèrent bonne (classe 3*)

La session d'évaluation de la dernière version de Semiabong, encore en production aujourd'hui, a été menée sur deux semaines, du 15 au 30 septembre 2016. À ce moment-là, il s'agit d'une version primaire du système qui n'intègre pas la fonction finale optimisant

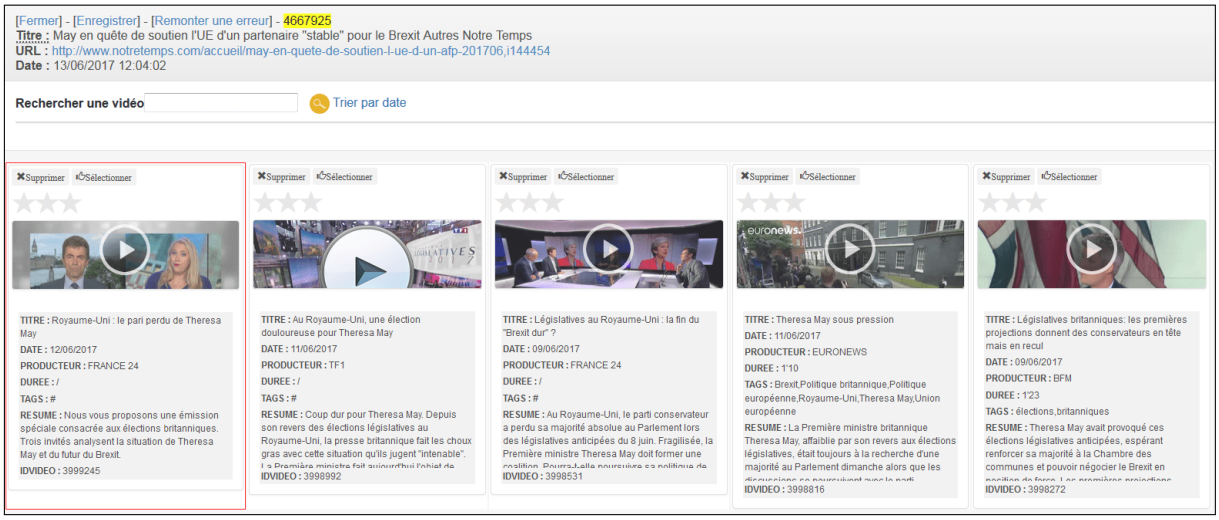

FIGURE 4.5 – Illustration de l'interface d'évaluation des vidéos pour les articles

la combinaison des deux scores associés à une paire article-vidéo (*score_thematique* et *score_date*). Quatre juges composent l'ensemble d'utilisateurs en charge de l'évaluation, ce qui nous permet d'obtenir pour cette session 860 jugements de pertinence, répartis tels que :

— classe 3* : 247 instances (soit 28.7%)
— classe 2* : 412 instances (soit 47.9%)
— classe 1* : 201 instances (soit 23.4%)

La classe des résultats *moyens* (*i.e.* classe 2*) est nettement sur-représentée, fait peu surprenant puisqu'elle constitue pour les juges la classe *fourre-tout* pour laquelle ils optent généralement. Les instances qui la composent présentent par ailleurs des valeurs de scores très hétérogènes, comme en témoigne la Figure 4.6 décrivant l'espace dans lequel on projette chaque paire article-vidéo en fonction de ses *score_thematique* et *score_date*.

On observe que les données de la classe 2* (en orange) ne constituent pas un ensemble clairement défini, mais ont tendance à s'étaler dans tout l'espace. Certaines d'entre elles, dans la partie basse de l'espace où sont concentrées beaucoup d'instances de la classe 1*, semblent s'apparenter à des résultats *mauvais*. Parallèlement, on en retrouve aussi un certain nombre dans des portions de l'espace qui concentrent des instances de la classe 3*, suggérant qu'il pourrait s'agir de *bons* résultats.

Cette forte dispersion des données de la classe 2*, ainsi que leur sur-représentation dans le jeu de données, a rapidement fait émerger le risque de générer un modèle sur-catégorisant les instances dans cette classe, s'il était appris sur cet ensemble de données. Par ailleurs, rappelons que notre objectif initial est de distinguer *strictement* les bons résultats des mauvais, en construisant un modèle capable de séparer deux ensembles de données. Considérer cette classe 2* dans l'apprentissage du modèle aurait donc peu

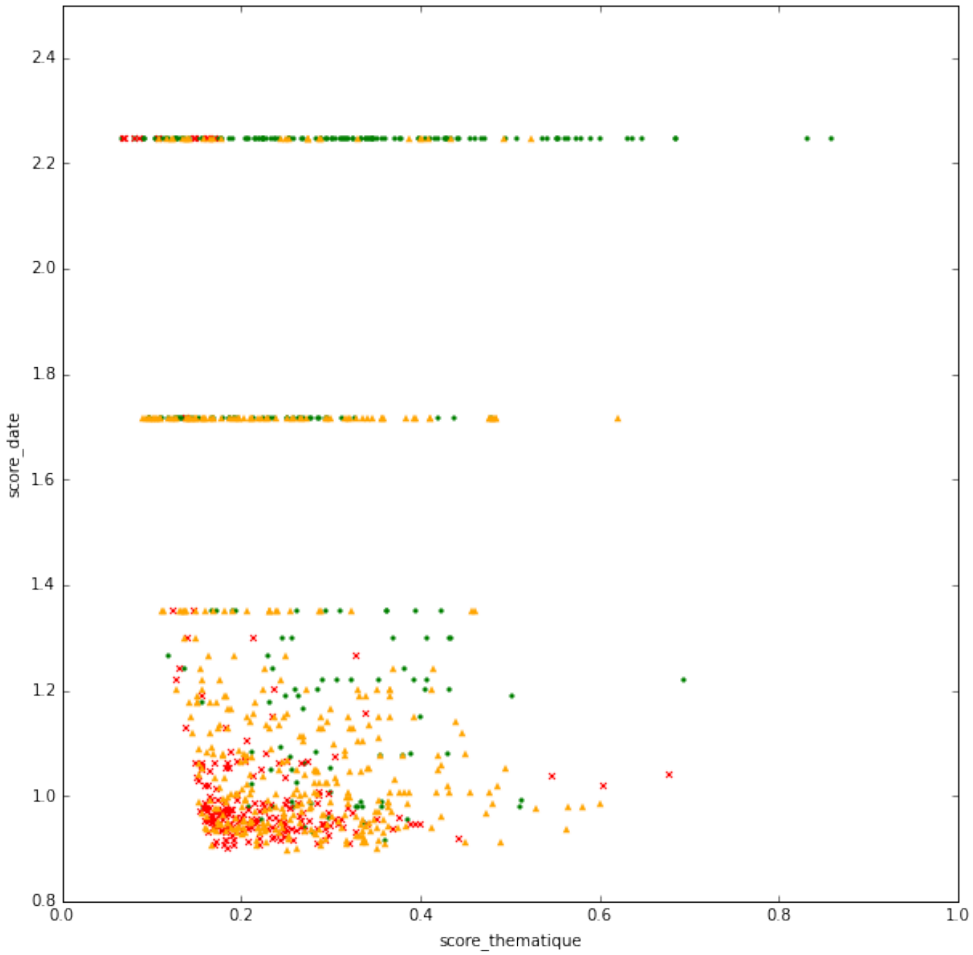

FIGURE 4.6 – Distribution des résultats bons (en vert), moyens (en orange) et mauvais (en rouge) dans le corpus annoté

d'intérêt dans cette perspective, puisque nous souhaitons au contraire pouvoir déterminer, parmi les données qui la composent, lesquelles devraient être rattachées à la classe 1* et lesquelles à la classe 3*.

En résumé, bien qu'un niveau intermédiaire d'annotation ait été nécessaire pour fiabiliser les jugements des classes *bon* et *mauvais*, permettant aux juges de n'attribuer ces notes extrêmes que lorsqu'ils le souhaitaient vraiment, les données composant la classe 2* ne seront pas exploitées lors de l'apprentissage du modèle de classification binaire. Une fois construit, ce modèle pourra distinguer, parmi ces appariements *moyens*, ceux qui s'approchent des *mauvais* de ceux qui s'approchent des *bons*.

4.6 Ordonnancement par classification

L'objectif ici est de construire un modèle capable d'ordonnancer les résultats du système en fonction des *score_thematique* et *score_date* afin de proposer en tête de clas-

sement celui offrant le meilleur compromis entre les deux. Par ailleurs, nous souhaitons pouvoir distinguer les bons résultats des mauvais afin d'éviter de présenter ces derniers aux utilisateurs. Considéré ainsi, le problème nous est apparu plus clairement comme un problème de classification binaire, pour lequel doivent être distingués deux ensembles de données dans un espace bidimensionnel. Les données exploitées pour cet apprentissage supervisé sont celles des deux classes 1* et 3* précédemment présentées, décrites par seulement deux traits que sont leurs valeurs de *score_thematique* et *score_date*.

Au regard de la configuration des données et du type de tâche à remplir, à savoir une discrimination, nous avons fait le choix d'implémenter un modèle SVM dont l'hyperplan calculé devrait optimiser la séparation des deux classes de données. Si d'autres modèles, tels qu'une régression linéaire ou logistique binaire, auraient également pu convenir pour une telle tâche, le choix d'un SVM s'est imposé pragmatiquement, en accord avec le reste l'équipe R&D de l'entreprise au sein de laquelle nous étions tous familiarisés avec les modèles SVM. Aussi, aucune expérience comparative n'a été menée pour déterminer le meilleur type de modèle à implémenter pour cette tâche de discrimination. Toutefois, parmi les différents modèles SVM, nous avons testé trois implémentations : une linéaire et deux polynomiales (de degrés 2 et 3). Les données des deux classes n'étant pas tout à fait linéairement séparables (bien que certains paquets se distinguent nettement d'un côté et de l'autre de l'espace tel que l'on peut l'observer en Figure 4.7), nous supposions qu'un SVM à noyau polynomial offrirait de meilleures performances. Or en comparant les résultats des différents modèles (en validation croisée à 10 plis) le SVM linéaire atteint une exactitude[10] moyenne de 0.85, égale à celle obtenue par un SVM polynomial de degré 3 et meilleure que celle obtenue par un SVM polynomial de degré 2 qui n'atteint que 0.84. En accordant un poids supplémentaire à la classe des bons résultats lors de l'apprentissage, ces valeurs d'exactitude augmentent pour chacun des modèles : 0.88 pour le SVM linéaire ; 0.87 pour les deux modèles à noyau polynomial. Nous conservons donc le modèle linéaire dans le système final, dont on peut observer l'hyperplan calculé sur les données d'apprentissage sur la Figure 4.7. Nous constatons que les bons appariements se concentrent majoritairement aux valeurs de *score_date* correspondant à une production vidéo datant du jour de la publication de l'article (*score_date* = 2.24) ou de la veille (*score_date* = 1.72). Parallèlement, la majorité des mauvais appariements sont concentrés dans une portion de l'espace, que l'hyperplan tracé – calculé par le SVM – permet en partie d'isoler.

À partir de ce modèle, il est possible d'établir un certain seuil distinguant les bons résultats des mauvais. Dorénavant, plutôt que d'implémenter le produit des *score_thematique* et *score_date* comme score final pour une paire article-vidéo, nous projetons cette paire dans un espace bidimensionnel tel que décrit en Figure 4.7 : si la paire formée par les

10. *cf.* Chapitre 1 pour une description du calcul de cette mesure.

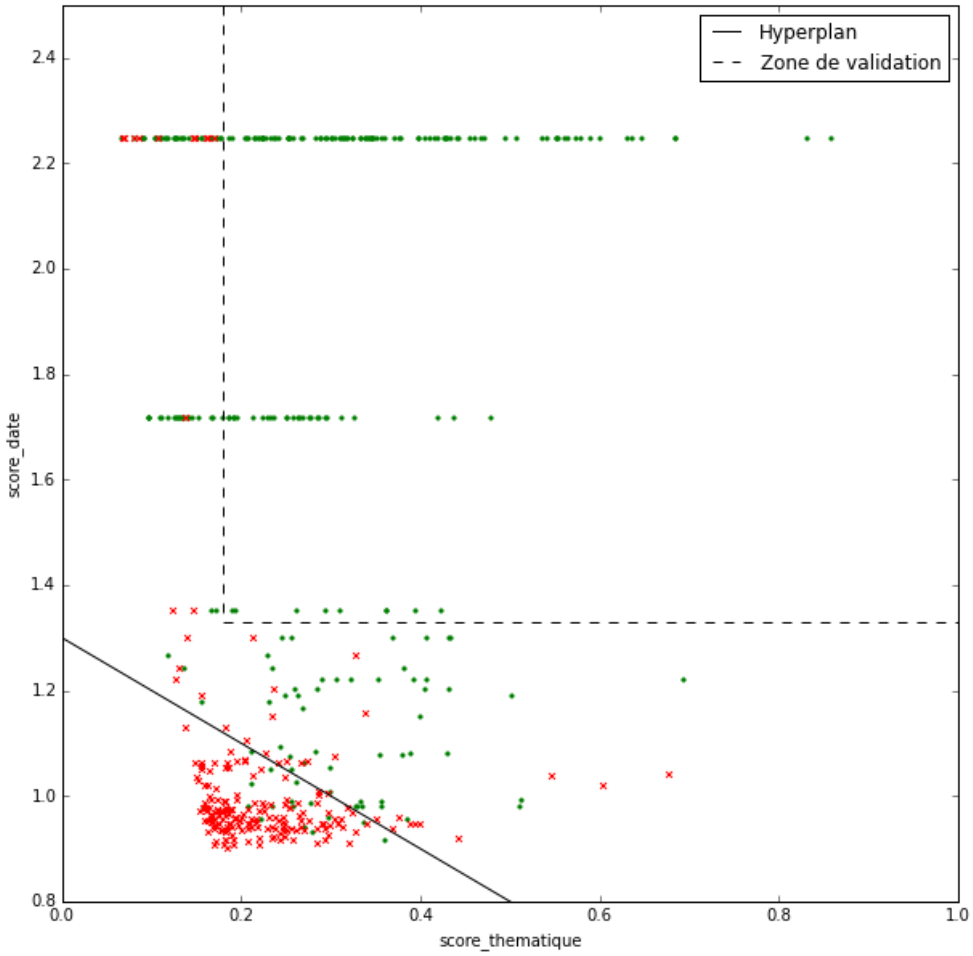

FIGURE 4.7 – Répartition des instances positives (en vert) et négatives (en rouge), en fonction des *score_thematique* et *score_date*

deux documents se trouve sous le séparateur, la vidéo est considérée comme un mauvais résultat pour l'article ; en revanche si elle est au-dessus, elle est considérée comme un bon résultat. En outre, ces bons résultats sont ordonnés en fonction de la distance séparant leurs représentations ponctuelles de l'hyperplan tracé : une vidéo est d'autant plus pertinente pour un article que la paire qu'ils forment est éloignée du séparateur [11].

Le séparateur calculé correspond à la droite D de la fonction affine de type $f(x) = a(x)+b$ où $a = -1$ et $b = 1.3$, telle que décrite ci-dessous en 4.6. Quant à la distance séparant un point A de coordonnées (x_A, y_A) de cette droite D, elle s'obtient par le calcul décrit en 4.7.

$$f(x) = -x + 1.3 \tag{4.6}$$

11. La meilleure vidéo possible étant celle de coordonnées (*1, 2.24*)

$$d(A,D) = \frac{|ax_A - y_A + b|}{\sqrt{1+a^2}} = \frac{|-x_A - y_A + 1.3|}{\sqrt{2}} \tag{4.7}$$

L'objectif final du système étant de minimiser l'intervention humaine dans l'appariement de vidéos aux articles, nous cherchons à déterminer un seuil au-delà duquel il est possible de valider automatiquement les paires de documents, sans la soumettre à une vérification manuelle. La Figure 4.7 présente une zone (délimitée par les pointillés) dans laquelle les appariements ont tous été jugés corrects par les évaluateurs, donc toute vidéo dont la paire formée avec un article se trouve dans cette zone est désormais considérée comme bonne et directement intégrée en bas d'article.

L'Annexe F présente un échantillon d'appariements article-vidéo automatiquement validés par le système, c'est-à-dire des cas où la vidéo sélectionnée est directement intégrée à l'article, sans vérification manuelle.

Notons qu'à partir de ce modèle, les instances de la classe 2* de notre corpus annoté, que nous avions exclu de notre ensemble d'apprentissage, se répartissent entre les deux classes, tel qu'illustré en Figure 4.8. Certains des appariements de cette classe se retrouvent même dans la zone de validation automatique, c'est-à-dire qu'en conditions réelles d'applications, les vidéos composant ces paires auraient directement été intégrées aux pages d'articles correspondant.

Le modèle calculé à partir des jugements de pertinence récupérés permet donc d'opérer trois tris parmi les résultats candidats, comme l'illustre la Figure 4.1. Dans un premier temps, les vidéos sont triées dans l'ordre décroissant de leur score final puis certaines sont exclues des résultats en fonction de leur position dans l'espace. Ainsi, il est possible qu'un article n'ait finalement aucun résultat retourné par le système, auquel cas il est présenté dans l'interface utilisateur sans aucun contenu associé. Les utilisateurs opèrent alors une recherche manuelle, qui peut elle-même aboutir à un résultat vide, si aucune des vidéos de la collection ne les satisfait. Dans le cas contraire où l'ensemble de résultats contient au moins une vidéo à l'issue du filtre sur les scores, une seconde condition s'applique, vérifiant si le score d'une vidéo lui permet d'être automatiquement validée par le système sans aucune vérification manuelle. Si tel est le cas, elle est alors directement associée à l'article. Par ailleurs, si plusieurs vidéos remplissent cette condition, c'est celle ayant obtenu le score le plus haut (*i.e.* la plus grande distance à l'hyperplan) qui est associée à l'article. Si au contraire aucune vidéo ne remplit cette condition, alors les cinq ayant obtenu les plus hauts scores sont présentées aux utilisateurs pour une vérification manuelle.

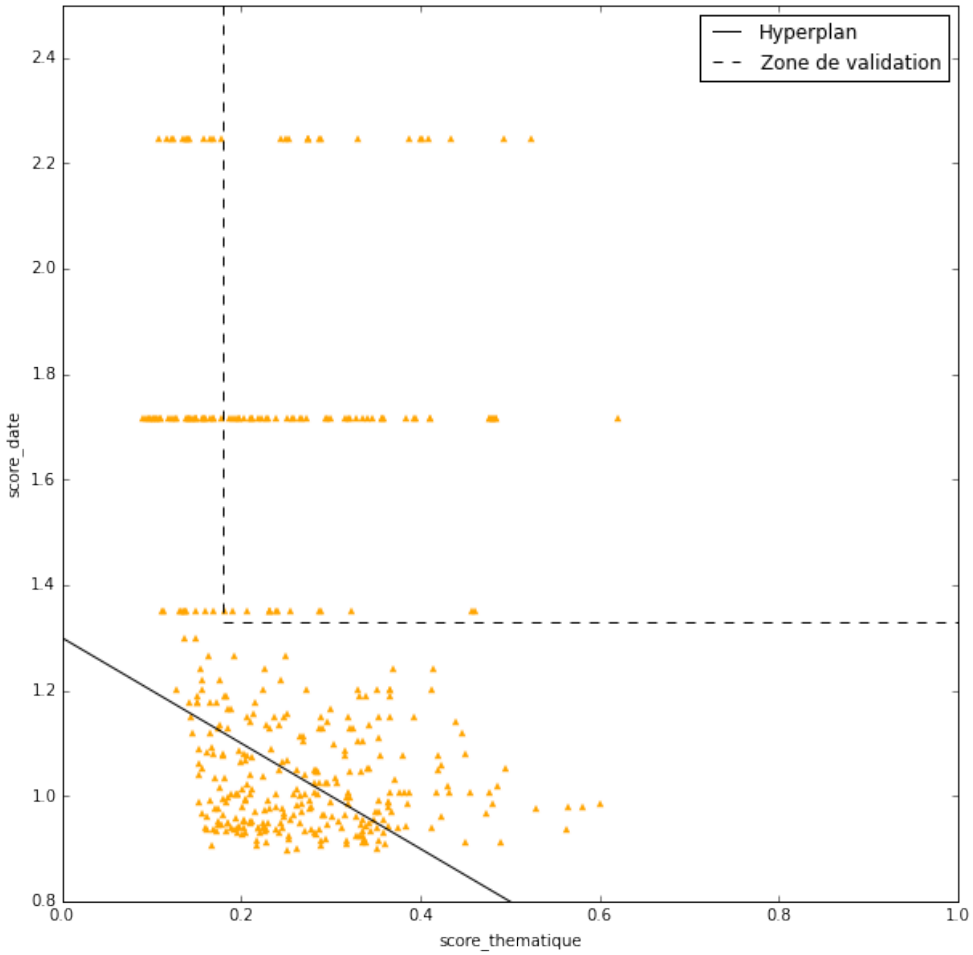

FIGURE 4.8 – Classification des données de la classe 2* du corpus annoté

4.7 Conclusion

Différents systèmes d'appariement d'articles en ligne et de vidéos ont été testés tout au long de cette thèse, et ce chapitre a présenté la chaîne de traitement de la dernière version développée, SEMIABONG, actuellement en production chez MEDIABONG.

Nous y avons exposé la façon dont les documents traités, articles et vidéos, sont indexés pour que le système en ait une représentation plus précise qu'un simple *sac de mots*. Cette représentation enrichie s'appuie notamment sur l'interrogation d'une ressource propriétaire dont les EMM référentielles qui la composent sont intégrées semi-automatiquement. Elle pourrait être davantage affinée en interrogeant des ressources externes telles que des bases de connaissance ou en faisant appel à des outils d'annotations linguistiques de différents niveaux (syntaxique et sémantique notamment) pour une analyse plus fine des textes. Néanmoins, afin de répondre efficacement aux contraintes industrielles de ce projet, les traitements ici implémentés tiennent plus de la fouille de textes, qui se focalisent essentiellement sur les données à disposition, que du TAL impliquant plus d'analyse linguistique.

Nous avons par ailleurs proposé, en plus du classique cosinus mesurant la similarité thématique entre deux représentations vectorielles de document, une fonction saisissant leur proximité temporelle. Nous avons en effet observé que dans le cadre du traitement de contenus d'actualité, la fraîcheur du résultat retourné pouvait s'avérer tout aussi importante que son degré de similarité thématique.

Enfin, nous avons présenté la façon dont les différents scores décrivant la similarité entre un article et une vidéo étaient exploités comme attributs décrivant les instances d'apprentissage d'un modèle d'ordonnancement des résultats. Cet apprentissage supervisé repose sur des exemples annotés manuellement par les utilisateurs du système dont on récupère les jugements de pertinence lors de sessions d'évaluation. L'ensemble d'attributs décrivant les paires article-vidéo, se limitant ici aux deux informations de scores, pourrait être élargi en intégrant par exemple des données sur les thématiques des documents, ou sur leurs producteurs. Le modèle retenu, un SVM linéaire, pourrait quant à lui confronter ses performances à d'autres fonctions de combinaison de scores, notamment certaines décrites dans l'état de l'art du *Topic Detection and Tracking* présenté au Chapitre 2.

Le chapitre suivant présente le protocole mis en place pour l'évaluation des différentes versions développées au cours de ces années de thèse et pour la comparaison de leurs performances respectives dans notre contexte particulier.

Chapitre 5

Évaluation des performances

L'évaluation des performances d'un système automatique est fortement dépendante des données qu'il traite et de la tâche qu'il doit accomplir. Dans notre contexte particulier, cette tâche est complexe puisqu'elle doit satisfaire différents besoins loin d'être complémentaires. Il nous a fallu trouver un compromis satisfaisant à la fois le besoin d'information exprimé par le contenu d'un article et le besoin d'intégration vidéo essentiel pour Mediabong d'un point de vue économique.

Il est rapidement apparu que les métriques classiques en RI n'étaient pas adaptées à notre problématique. Nous proposons donc ici un protocole d'évaluation des performances du système développé considérant l'ensemble des contraintes qui sont les nôtres.

Nous commençons dans ce chapitre par exposer ces contraintes, en les mettant en perspective avec les protocoles traditionnels de la RI (Section 5.1). Puis nous présentons en Section 5.2 la façon dont nous avons adapté l'évaluation des performances à notre cadre industriel. Nous revenons notamment sur la façon dont y sont comparés différents systèmes entre eux (Section 5.2.1), ainsi que sur les métriques implémentées pour rendre compte de leurs capacités à remplir les objectifs visés (Section 5.2.2). Avant de conclure, nous décrivons en Section 5.3 les performances du système actuellement en production depuis la fin du mois de décembre 2016, satisfaisant les besoins initiaux formulés par Mediabong.

5.1 Spécificités du contexte

Par rapport à un SRI dit *classique*, notre étude présente plusieurs spécificités liées aux contenus analysés et à la tâche à automatiser. Cette section illustre ces particularités par des exemples extraits du corpus puis expose les contraintes qu'elles engendrent pour l'évaluation des performances du système.

5.1.1 Exemples

Cette section présente un éventail d'exemples comparant les vidéos proposées automatiquement pour un article à celles sélectionnées manuellement par un des utilisateurs du système. Dans chacun de ces exemples, on note **Vidéo auto** la vidéo proposée par le système et **Vidéo manuelle** la vidéo sélectionnée par l'utilisateur pour former une paire avec l'**Article**.

En premier lieu, la collection de vidéos est à la fois non-exhaustive et dynamique. Non-exhaustive car un article soumis au système n'a pas nécessairement de vidéo associable en base et dynamique car il peut en avoir qui ne sont pas encore indexées ni même encore produites au moment du traitement automatique. C'est notamment le cas pour les articles relatant des faits-divers locaux dont la spécificité est telle qu'aucun des partenaires producteurs n'a de vidéo sur le sujet – *cf.* **(1)**. On trouve également parmi ces cas particuliers des articles relatant des sujets très spécifiques – *cf.* **(2)**, **(3)** – non illustrés dans la collection vidéo. Il s'agit donc pour le système de ne rien proposer en réponse à ces articles, plutôt que de retourner une ou plusieurs vidéos sans aucun lien avec leur contenu. Les exemples cités ci-dessus correspondent à des cas d'articles pour lesquels ni le système ni l'utilisateur n'ont associé de vidéo.

(1)			
Article	(2016-21-11)	-	*Picardie : Un chasseur tué d'un ricochet en pleine tête*
Vidéo auto	∅	–	∅
Vidéo manuelle	∅	–	∅
(2)			
Article	(2016-09-01)	–	*Robert Neuburger : Accusez-vous le sort ou bien pensez-vous que vous y êtes pour quelque chose ?*
Vidéo auto	∅	–	∅
Vidéo manuelle	∅	–	∅
(3)			
Article	(2016-11-18)	–	*Une recherche Google, c'est combien de C02 ?*
Vidéo auto	∅	–	∅
Vidéo manuelle	∅	–	∅

Par ailleurs, la collection de vidéos étant riche, certains articles sont susceptibles d'y trouver plus d'un résultat satisfaisant. C'est particulièrement le cas des articles relatant des faits d'actualités, majoritaires dans nos données. Beaucoup de diffuseurs et producteurs proposent en effet des contenus en direct et en temps réel qui suivent l'actualité

médiatique. De ce fait, lorsqu'un fait marquant survient à un moment donné, beaucoup de contenus – articles comme vidéos – relatent ce sujet. Or nous souhaitons n'associer qu'une unique vidéo à l'article. Les exemples **(4)**, **(5)** et **(6)** présentent trois cas d'articles pour lesquels le système propose un résultat que l'utilisateur rejette au profit d'un autre non proposé automatiquement. Ils illustrent la difficulté de saisir, même manuellement, les critères sur lesquels se basent les utilisateurs pour sélectionner une vidéo pour un article.

(4)			
Article	(2016-09-01)	–	*L'Allemand Thomas Buberl prend la tête du groupe*
Vidéo auto	(2016-06-21)	–	*Le nouveau patron d'Axa dévoile sa stratégie*
Vidéo manuelle	(2016-06-01)	–	*Le regard de Challenges : Nicolas Moreau démissionne d'Axa*
(5)			
Article	(2016-11-18)	–	*L'activité physique bénéfique pour les patients atteints de Parkinson*
Vidéo auto	(2016-08-26)	–	*Maladie de Parkinson : une maladie du mouvement*
Vidéo manuelle	(2016-09-07)	–	*La maladie de Parkinson détectée grâce à vos yeux*
(6)			
Article	(2017-03-13)	–	*Chiens renifleurs de cancer : une efficacité à 100% sur six mois de tests*
Vidéo auto	(2016-08-27)	–	*KDOG, le projet qui utilise l'odorat du chien pour détecter le cancer du sein*
Vidéo manuelle	(2016-04-13)	–	*Et si les chiens pouvaient détecter les cancers ?*

En **(4)**, le résultat proposé par le système est au sujet du *nouveau patron d'Axa*, qui n'est autre que *Thomas Buberl*, sujet du titre de l'article. De plus, cette vidéo automatique est plus récente, par rapport à la date de publication de l'article, que la vidéo manuellement sélectionnée. Pourquoi alors décider de l'exclure en faveur d'une autre plus ancienne, qui plus est, au sujet de l'ancien patron d'Axa, *Nicolas Moreau* ?

En **(5)**, la vidéo proposée automatiquement et celle manuellement sélectionnée nous semblent aussi pertinente l'une que l'autre. Aucune des deux n'est l'exacte illustration du contenu de l'article mais elles abordent malgré tout le même sujet, la maladie de Parkinson. On suppose alors que c'est le paramètre temporel qui a influencé le choix de l'utilisateur, préférant sélectionner une vidéo plus récente que celle que proposait le système. Toutefois, l'article de cet exemple relate un contenu que l'on considère dans ces travaux comme froid, atemporel. On s'interroge alors sur l'intérêt porté à ce paramètre dans ce cas particulier. Par ailleurs, l'exemple **(6)** présente le cas inverse où l'utilisateur préfère sélectionner une vidéo plus ancienne que celle proposée par le système, alors que celle-ci répond selon nous correctement à l'article.

En parallèle de cela, nous évoquions plus haut le fait que le besoin initial formulé par MEDIABONG dépasse le seul besoin d'information, puisque son modèle économique est basé sur le nombre de vues de vidéos, donc sur le nombre d'articles ayant une vidéo associée. Il nous faut donc trouver un compromis satisfaisant chacune de toutes les exigences, en recherchant la meilleure vidéo pour un article si elle existe, sans toutefois rejeter une vidéo moyennement pertinente que nous préférons voir intégrée à un article plutôt que de ne rien lui associer. Pour illustrer cette difficulté, l'exemple **(7)** présente un cas d'article pour lequel l'utilisateur a sélectionné une vidéo tandis que le système n'avait rien proposé ; l'exemple **(8)** correspond à l'inverse au cas où le système propose un contenu que l'utilisateur rejette pour préférer ne rien sélectionner. Si nous concédons qu'en **(8)**, la proposition automatique n'est pas optimale, la sélection manuelle ne semble pas plus à-propos en **(7)**. On suppose ici que le fait qu'elle ait été produite le même jour a fait de cette vidéo un résultat convenable pour l'article. Mais nous notions plus haut que l'intérêt porté à ce paramètre temporel n'était pas constant chez les juges.

(7)

Article	(2016-10-06)	–	*Vladimir Poutine rencontrera François Hollande en France le 19 octobre*
Vidéo auto	∅	–	∅
Vidéo manuelle	(2016-10-06)	–	*Moscou restreint sa coopération nucléaire avec Washington*

(8)

Article	(2016-10-04)	–	*Syrie : les Etats-Unis n'ont pas renoncé à rechercher la paix*
Vidéo auto	(2016-09-30)	–	*Kerry : "Nous sommes en passe de suspendre les discussions" avec la Russie*
Vidéo manuelle	∅	–	∅

Tous ces exemples témoignent de la difficulté d'exploiter les résultats manuels attribués par des utilisateurs dont les sélections varient selon des critères difficiles à cerner. En l'absence de guide d'annotation clair établissant des règles strictes quant à l'attribution de jugements, on observe de fortes variations dans ces retours d'utilisateur, empreints de subjectivité. Cette instabilité introduit un fort biais dans les données, qui sont celles exploitées par les algorithmes d'apprentissage cherchant des régularités pour modéliser le besoin de ces utilisateurs.

Nous présentons plus loin dans ce chapitre (en 5.4) une mesure de l'accord entre les différents utilisateurs participant à l'évaluation des résultats de SEMIABONG, afin de rendre compte du degré de variation séparant leurs jugements.

5.1.2 Contraintes méthodologiques

Traditionnellement, les résultats d'un SRI sont évalués par le paradigme de Cranfield (Cleverdon, 1967) qui reproduit en conditions expérimentales un processus de RI. Cette méthode nécessite de sélectionner *a priori* un ensemble de requêtes auxquelles des experts associent l'ensemble des documents pertinents dans la collection considérée, pour former un ensemble de test. Les sorties du système automatique peuvent ensuite être comparées aux résultats de référence pour mesurer ses performances.

La définition-même de ce paradigme suppose que, dans un ensemble de test, toute requête a au moins un résultat dans une collection de documents fixe. Elle suppose également que nous connaissons et souhaitons retrouver tous les documents pertinents pour une requête.

Or parmi ces conditions, aucune n'est remplie dans le contexte de notre étude. En effet, un article soumis au moteur n'a pas systématiquement de vidéo associable en base. Quant à ceux qui en ont, il est inenvisageable de rechercher manuellement tous ceux qui sont pertinents, tant la collection est grande et les juges peu disponibles.

Par ailleurs, même si un tel ensemble était construit pour un article à un instant t, il serait potentiellement obsolète à $t+1$ du fait de l'évolution permanente de la collection vidéo à laquelle sont quotidiennement ajoutés de nouveaux documents et régulièrement supprimés de trop anciens.

Finalement, le système développé n'a pas vocation à retourner exhaustivement les résultats pertinents en réponse à une requête. Sa tâche est une tâche d'appariement pour laquelle nous souhaitons optimiser autant que possible la précision au premier document ($P@1$), bien que la non-exhaustivité de la collection de vidéos ne le permette pas toujours.

Ce cadre strict a en fait évolué avec les travaux en RI sur le Web, comme (Garg et al., 2012), dont les contraintes excluaient initialement une évaluation dans ce paradigme. Ils proposent donc de ne travailler que sur une portion figée de la collection de documents (de plusieurs millions d'entrées pour rester représentatif de la collection initiale) pour pallier la contrainte de dynamisme du Web. Cependant, la tâche d'attribution de jugement de pertinence reste un problème en l'état puisque nous ne pouvons imaginer qu'un expert analyse la pertinence des quelques millions de documents pour chacune des requêtes. (Garg et al., 2012) font ici appel à des méthodes de *pooling*, permettant de sélectionner un sous-ensemble de résultats pour chacune des requêtes, en fusionnant les ensembles de n meilleurs documents retournés par différents SRI. Ainsi, les experts n'ont plus qu'à juger la pertinence des documents de ce sous-ensemble pour chaque requête, le reste des documents de la collection étant par ailleurs considérés comme non-pertinents. Si cette méthode permet de gagner un temps considérable en annotation de l'ensemble de test,

elle doit être correctement implémentée, notamment dans le choix de la variable n, pour ne pas biaiser les résultats de l'évaluation (Buckley et al., 2007).

Bien que ce cadre considère mieux que le paradigme initial les contraintes qui sont les nôtres, il reste expérimental et difficilement transposable à nos travaux qui doivent satisfaire plus qu'un besoin d'information, en conditions réelles d'application. Dans notre contexte, un résultat n'est pas intrinsèquement pertinent pour une requête, il l'est par rapport au reste de la collection, et par rapport à la tâche elle-même. Que s'ajoute au besoin d'information un besoin d'optimisation d'intégration vidéo dans les pages fait qu'une vidéo pourrait être jugée non pertinente en réponse au besoin d'information pour un article mais être malgré tout sélectionnée par l'utilisateur afin de satisfaire le besoin d'intégration. Du même fait, ils nous est difficile d'envisager de figer la collection de documents pour l'ensemble de test, car l'attribution ou non d'une vidéo à un article dépend fortement de la disponibilité de résultats en collection au moment-même du traitement de l'article.

Face à ces contraintes, il a fallu envisager une procédure d'évaluation personnalisée, rendant compte des performances du système relativement aux exigences particulières à satisfaire. La section suivante en détaille les choix et leurs applications.

5.2 Adaptation du protocole

Cette section détaille la façon dont nous proposons de comparer différents systèmes dans notre contexte particulier où aucun ensemble de test n'a été créé. Il s'agit de rendre compte des performances globales du système, relativement à toutes les exigences établies, et non seulement sa capacité à répondre correctement au besoin d'information.

5.2.1 A/B testing

Les résultats du système d'appariement sont extrêmement sensibles aux données traitées en requête puisque les sujets abordés dans les articles et vidéos suivent l'actualité. Si un événement important survient un jour précis, alors la plupart des articles publiés ce jour-ci traiteront de cet événement et parallèlement beaucoup de vidéos seront produites à ce sujet. En revanche, si aucun événement particulier ne survient au cours d'une journée, nous observons une plus forte diversité des sujets traités dans les articles et dans les vidéos produites. Un système aura donc plus de chances d'atteindre de bonnes performances sur le premier jour que sur le second. C'est pourquoi construire un ensemble de test statique ne saurait rendre compte de la dépendance des performances aux sujets traités en temps réel.

Cette variabilité de résultats en fonction de l'actualité est prise en compte lors de l'éva-

luation des performances du système en considérant un intervalle de temps suffisamment long pour être représentatif de la diversité des contenus traités : nous faisons ici le choix d'un mois. Suivant la même idée, pour comparer deux versions du système, nous les poussons en production simultanément, sur un mode d'A/B testing, pour qu'elles traitent des données comparables, faute de données identiques. Nous comparons à la fin du mois les performances respectives de chacune des deux pour sélectionner la meilleure, en fonction des métriques proposées par la suite.

Contrairement à l'évaluation des données destinées à l'apprentissage du modèle d'ordonnancement, il s'agit ici d'une évaluation implicite pour laquelle aucun jugement de pertinence n'est attribué. Les performances du système sont déduites de la comparaison des résultats proposés automatiquement et de ceux sélectionnés manuellement par les utilisateurs.

Dans cette configuration, le système propose pour chaque article traité entre zéro et cinq résultats aux utilisateurs. Celui en charge de la vérification d'un article particulier peut soit sélectionner une des propositions du système, si toutefois il y en a une ; soit sélectionner un résultat hors de ces propositions ; soit ne rien sélectionner du tout s'il estime qu'aucun contenu de la collection ne répond correctement au contenu de l'article.

Les deux versions A et B sont représentées de façon quasi-équivalente dans ces données. Il nous est difficile de garantir une égalité stricte entre les deux bien qu'elles tournent simultanément en production, puisque nous n'avons pas la main sur ce que traitent les utilisateurs en cours d'évaluation. Contrairement à de l'A/B testing standard, nous ne divisions pas l'effectif des utilisateurs en deux, pour que l'un évalue la version A tandis que l'autre évalue la B. Tous ont accès à tous les articles soumis au système et traitent tous ceux qui s'offrent à eux lors de la session d'évaluation.

Notons enfin que lors de ces sessions, le module de validation automatique des vidéos sans vérification manuelle, présenté en fin de chapitre 4, est désactivé. Il permet en effet d'intégrer directement une vidéo à un article si son score de similarité dépasse un certain seuil. En conservant cette condition, on se priverait potentiellement pour l'évaluation de quantité de données illustrant les cas où l'utilisateur sélectionne la première vidéo proposée par le système. Or il nous intéresse particulièrement de savoir dans quelles proportions l'utilisateur apprécie ce premier résultat, nous renseignant sur la qualité de la fonction d'ordonnancement. Le modèle de classification sur lequel se base l'ordonnancement des résultats est, du reste, dépendant de chacune des versions. Si la fonction de *score_date* reste la même pour chacune d'elle, les valeurs de *score_thematique* varient de l'une à l'autre. Or le calcul du séparateur se base sur ces deux paramètres pour construire l'hyperplan optimal, il est donc dépendant de la version du système traitant les données exploitées pour l'apprentissage.

5.2.2 Tâche de classification

Le fait que tous les articles n'aient pas de résultats en base nous fait reconsidérer le problème comme une tâche de classification binaire, où les deux classes sont celle des articles ayant des résultats en base (notée VIDEO), et celle des articles n'ayant aucun résultats en base (notée NO VIDEO). La répartition des données selon ces deux classes nous renseigne sur les performances du système en réponse au besoin de maximisation d'intégration vidéo, en nous permettant d'observer les cas où utilisateur et système sont d'accord sur le fait de sélectionner ou non une vidéo pour un article et inversement les cas où ils sont en désaccord.

Pour mesurer parallèlement les performances en réponse au besoin d'information, les éléments de la classe VIDEO sont ensuite subdivisés en différents cas selon le niveau de correspondance entre les résultats automatiques et la sélection manuelle : sont considérés comme très bons les cas où l'utilisateur sélectionne la première vidéo proposée par le système (noté *First*) ; comme moins bons les cas où il sélectionne l'une des autres du top 5 (noté *Top 5*) ; comme mauvais ceux où il sélectionne une vidéo qui n'a pas été proposée automatiquement (noté *Outside*).

La Table 5.1 présente une matrice de confusion typique représentant ces différents cas de figure. Nous pouvons ainsi calculer certaines métriques relatives à la tâche de classification, comme la précision et le rappel des deux classes. Une telle répartition nous permet également de mesurer la précision au premier document (*P@1*), qui nous intéresse particulièrement pour notre tâche d'appariement. Cette mesure ne s'intéresse qu'aux cas pour lesquels le système propose au moins un résultat et s'obtient ici directement en divisant le nombre de cas pour lesquels les utilisateurs ont sélectionné la première vidéo (*First*) à l'ensemble des cas pour lesquels le système propose au moins une vidéo.

<table>
<tr><th colspan="2">Reference / System</th><th>VIDEO</th><th></th><th>NO VIDEO</th></tr>
<tr><td rowspan="3">VIDEO</td><td>First</td><td>V_{SYS_f}</td><td rowspan="3">V_{ALL}</td><td rowspan="3">NV_{REF}</td></tr>
<tr><td>Top 5</td><td>V_{SYS_t}</td></tr>
<tr><td>Outside</td><td>V_{SYS_o}</td></tr>
<tr><td colspan="2">NO VIDEO</td><td colspan="2">NV_{SYS}</td><td>NV_{ALL}</td></tr>
</table>

TABLE 5.1 – Modèle générique de matrice de confusion

La mesure du rappel de la classe VIDEO (équation 5.3) et celle de la précision de la classe NO VIDEO (équation 5.2) nous renseignent sur la capacité d'intégration vidéo du système. De hautes valeurs sur ces métriques rendent compte de ses bonnes performances quant à sa capacité à trouver des résultats pour les articles traités. Ces valeurs ne fournissent en revanche aucun renseignement sur la qualité du résultat fourni. C'est en observant la précision au premier document (5.5) que l'on est en mesure de conclure sur la qualité de

la vidéo en tête de résultat, qui est celle que nous souhaitons optimiser. Par ailleurs, la précision de la classe VIDEO (équation 5.1) et parallèlement le rappel de la classe NO VIDEO (équation 5.4) nous informe sur la capacité du système à détecter les articles pour lesquels aucun résultat satisfaisant n'existe dans la collection vidéo. Elles représentent les cas où le système, comme l'utilisateur, a préféré n'associer aucun résultat à un article plutôt que de lui en associer un mauvais. Des valeurs élevées sur ces métriques rendent compte des bonnes performances quant à la capacité du système à filtrer les mauvais résultats. Elle est essentielle afin d'éviter d'intégrer aux articles des vidéos que les diffuseurs pourraient réprouver.

$$P_{Video} = \frac{V_{ALL}}{V_{ALL} + NV_{REF}} \tag{5.1}$$

$$P_{NoVideo} = \frac{NV_{ALL}}{NV_{ALL} + NV_{SYS}} \tag{5.2}$$

$$R_{Video} = \frac{V_{ALL}}{V_{ALL} + NV_{SYS}} \tag{5.3}$$

$$R_{NoVideo} = \frac{NV_{ALL}}{NV_{ALL} + NV_{REF}} \tag{5.4}$$

$$P@1 = \frac{V_{SYS_f}}{V_{ALL} + NV_{REF}} \tag{5.5}$$

La Table 5.2 présente la répartition des articles selon ces cas, pour les versions A (correspondant à l'actuelle version en production, dont les spécificités sont présentées au chapitre 4) et B (correspondant à la version antérieure développée), comparées sur le mois d'octobre 2016. La Table 5.3 présente les valeurs obtenues pour chacune des versions A et B selon chacune des métriques évoquées.

		Version A		Version B	
System \ Reference		VIDEO	NO VIDEO	VIDEO	NO VIDEO
VIDEO	*First*	433	342	153	142
	Top 5	10		4	
	Outside	60		72	
NO VIDEO		87	373	181	516

Table 5.2 – Matrices de confusion pour les versions A & B

Les résultats démontrent de grandes variations de performances entre ces deux versions. La version B en offre de bien meilleures pour détecter les articles sans résultats en

		Version A	Version B
Precision	Video	0.59	**0.61**
	No Video	**0.81**	0.74
Recall	Video	**0.85**	0.56
	No Video	0.52	**0.78**
P@1		**0.51**	0.32

Table 5.3 – Résultats de la classification pour les versions A & B

collection (scores élevés pour les métriques de P_{Video} et $R_{NoVideo}$), mais elle est en revanche bien moins efficace que la version A quant au nombre de vidéos proposées en réponse aux articles. Elle n'atteint en effet qu'un score de 0.56 pour le R_{Video} tandis que la version A atteint 0.85 sur cette métrique. Ces deux observations semblent indiquer que la version B a des seuils de score bien plus restrictifs que la A et a tendance à peu proposer de résultats dans la plupart des cas. À l'inverse, les résultats de la version A semblent indiquer que ses seuils de score sont un peu trop tolérants, puisqu'elle a tendance à majoritairement proposer des résultats en réponse aux articles, alors que certains d'entre eux auraient dû être exclus.

Ce constat démontre la forte dépendance des systèmes aux données utilisateurs exploitées lors de l'apprentissage d'ordonnancement présenté au chapitre 4. Ce qui distingue ici particulièrement les deux versions est la capacité à correctement classer un résultat comme bon ou mauvais en réponse à un article. La version B semble avoir tendance à privilégier la classification des instances en mauvais tandis que la A présente la tendance inverse. Les modèles appris pour chacune des versions dépendent des jugements récoltés lors des sessions d'évaluation qui ne se font pas sur les mêmes données d'une version à l'autre. En outre, ces données sont dépendantes de l'ensemble des utilisateurs du système chargés de l'évaluation, dont l'effectif et les attentes individuelles [1] peuvent également varier à chaque session. Ces problématiques relatives à l'évaluation des performances nous ont beaucoup interrogés tout au long de cette thèse, du fait de cette difficulté à stabiliser le protocole pour en fiabiliser les résultats. Il apparaît clairement, au regard de ceux présentés dans ce chapitre, que la méthode proposée ici présente quelques lacunes. Malgré tout, elle nous permet d'estimer les performances générales des versions testées et de faire un choix parmi celles-ci relativement aux objectifs visés.

L'objectif principal dans notre contexte industriel est de maximiser la précision au premier document. Rappelons en effet que la demande initiale de Mediabong était de réduire autant que possible l'intervention humaine dans la tâche d'appariement article-vidéo. Or c'est la première vidéo retournée par le système, celle ayant obtenu la plus haute valeur de *score_final*, qui est susceptible d'être automatiquement appariée. C'est

1. La section 5.4 détaille quelques exemples de désaccord entre annotateurs.

donc prioritairement la métrique de *P@1* qui nous permet de choisir une version plutôt qu'une autre. C'est ainsi que la comparaison des versions A et B présentée ici nous a permis de sélectionner la version A, en production stable depuis le mois de novembre 2016, puisqu'elle atteint une *P@1* de 0.51, contre seulement 0.32 pour la version B.

La comparaison de ces deux versions illustre la procédure d'évaluation mise en place chez MEDIABONG. Chaque nouvelle version développée est comparée à la précédente mise en production. Ces comparaisons ne se font donc toujours que deux à deux et l'on conserve à l'issue de chacune d'elles la version la plus performante qui sera confrontée à la version suivante.

Au cours de cette thèse, sept versions différentes ont été testées et comparées sur ce principe. Toutes se basent sur une approche vectorielle, mais se distinguent au niveau des représentations documentaires, notamment au niveau des pondérations des différents types de termes [2]. La version conservée à l'issue de chaque comparaison reste en production, tandis que l'autre en est retirée. Le module de validation automatique est ensuite réactivé, et l'on peut observer les semaines suivantes dans quelles proportions le système valide automatiquement les résultats. L'objectif initial fixé par MEDIABONG était d'être capable de traiter au moins 70% des articles de façon complètement automatisée, les utilisateurs n'ayant alors plus que 30% de l'ensemble à vérifier manuellement.

5.3 Performances globales de Semiabong

Depuis sa mise en production au mois de novembre 2016, le système SEMIABONG a démontré de bonnes performances à tous les niveaux d'exigence établis par MEDIABONG. Sur sept mois de traitement du 1er novembre 2016 au 30 mai 2017, il a traité 36 725 articles, soit en moyenne 187 articles par jour. Parmi eux, 26 760 – soit 72.9% – ont été traités intégralement par le système, sans aucune validation manuelle. Cette performance dépasse donc de presque 3% les attentes initiales de l'entreprise.

Les 9 965 articles restant se répartissent entre les différents cas présentés dans ce chapitre, et sont recensés dans la Table 5.4. Les métriques rendant compte des performances du système sont quant à elles exposées en Table 5.5.

Nous observons de manière flagrante que les résultats obtenus sur ces sept mois de traitement se distinguent de ceux obtenus lors de la session de comparaison des versions A et B. SEMIABONG correspond pourtant bien à cette version A, puisqu'aucune modification ne lui a été apportée depuis qu'elle a été poussée en production en novembre 2016.

Les données ne sont évidemment pas comparables, ni en termes de contenu, ni en

2. Le chapitre 6 présente une série d'expérimentations où l'on peut observer l'influence de ces différentes représentations documentaires sur les résultats de RI.

System \ Reference		VIDEO		NO VIDEO
VIDEO	*First*	1401	1786	475
	Top 5	56		
	Outside	329		
NO VIDEO		2915		4789

TABLE 5.4 – Répartition des cas traités par SEMIABONG et validés manuellement, entre novembre 2016 et mai 2017

termes de quantité. Les résultats obtenus ici, confrontés à ceux obtenus par SEMIABONG lors de la session d'A/B testing, démontrent toutefois une forte dépendance des performances du système aux données. Alors que les précédents résultats tendaient à démontrer une meilleure performance du système en termes d'intégration vidéo qu'en termes de détection d'articles sans résultats, les résultats globaux démontrent la tendance inverse. Avec de fortes valeurs sur les métriques de P_{Video} et de $R_{NoVideo}$, SEMIABONG s'avère globalement efficace pour préférer ne rien sélectionner plutôt que de proposer un mauvais résultat en réponse à un article. En revanche, la faible valeur obtenue pour la métrique R_{Video} souligne qu'il a tendance à filtrer trop de résultats. L'une des conclusions possibles au regard de ces deux observations est que le classifieur distinguant les bons résultats des mauvais à tendance à sur-classifier les instances en mauvais. Nous avions observé, sur les résultats de la session d'A/B testing, la tendance inverse pour ce même système.

		SEMIABONG
Precision	VIDEO	0.79
	NO VIDEO	0.62
Recall	VIDEO	0.38
	NO VIDEO	0.91
P@1		0.62

TABLE 5.5 – Résultats de SEMIABONG sur la période novembre 2016 – mai 2017

Témoin d'une bonne capacité d'ordonnancement, le bon résultat obtenu pour la *P@1* nous conforte dans l'idée que les appariements automatiques sont pertinents. Ces cas particuliers n'étant toutefois pas soumis à une vérification manuelle, il nous est impossible d'en évaluer l'accord entre système et utilisateur. Or ils représentent une grande majorité des cas dans les résultats de SEMIABONG, dont il est essentiel de mesurer la qualité pour rendre compte de ses performances globales.

Conscients de cette lacune dans le protocole proposé ici, nous présentons par la suite une évaluation explicite de ces cas d'appariements particuliers.

5.4 Accord inter-annotateurs

Tel que nous l'évoquions au chapitre 4, une vidéo est automatiquement associée à un article lorsque la paire qu'ils forment se trouve dans une zone particulière de l'espace bidimensionnel représentant leurs scores d'association. Or cette zone est délimitée sur la base de la répartition des données annotées disponibles lors de la phase d'apprentissage. Comme dans tout modèle appris, si les exemples fournis en entrée sont trop spécifiques, le modèle de classification est peu généralisable et donc peu performant sur de nouvelles données.

Les cas d'appariements automatiques représentant la majeure partie des résultats de SEMIABONG, nous avons souhaité nous assurer de leur pertinence. Une session d'évaluation explicite a donc été mise en place au cours du mois de novembre 2016, après sa mise en production, sur ces données particulières. Sur la même échelle de notation ternaire que celle utilisée pour l'annotation des données destinées à l'apprentissage du modèle [3], cette évaluation a mobilisé deux juges.

Le système ayant vocation à intégrer une unique vidéo pour un article, indépendamment de l'utilisateur, l'annotation des données destinées à l'apprentissage du modèle de classification ne distinguait pas les juges entre eux. Il est toutefois apparu par la suite que les différents utilisateurs mobilisés pour l'évaluation ne s'entendaient pas toujours sur l'appréciation d'un résultat. Aussi, afin de mesurer la proportion de l'accord entre les différents juges, nous avons mené cette session d'évaluation explicite en double aveugle, *i.e.* en soumettant aux deux juges le même ensemble de données.

Le fait que l'appariement d'une vidéo à un article ait à satisfaire un besoin d'intégration vidéo en plus d'un besoin d'information peut biaiser le jugement par une sur-appréciation des résultats proposés par rapport à leur pertinence réelle. Par ailleurs, le fait qu'il s'agisse d'une tâche d'appariement, pour laquelle une seule vidéo est finalement associée à un article, tend au contraire à une sous-appréciation des résultats par des juges toujours en quête de la meilleure vidéo disponible possible. Ces deux comportements ont été observés respectivement chez les deux juges sollicités pour l'évaluation en double aveugle, chacun appréciant librement les appariements proposés selon ses propres critères. Ainsi, en leur soumettant un ensemble de 189 paires article-vidéo calculées et automatiquement validées par SEMIABONG, nous obtenons la répartition de jugements présentée en Table 5.6.

L'utilisateur 1 apparaît nettement moins satisfait que son pair, avec seulement 29.6% de bons jugements de pertinence (3*) contre 61.9% pour l'utilisateur 2. Sur les 117 cas que ce dernier a jugé pertinents, l'utilisateur 1 a majoritairement préféré nuancer sa note en n'attribuant qu'un jugement moyen (2*), révélant deux stratégies d'appréciation bien

3. C'est-à-dire l'attribution d'une étoile pour un appariement jugé mauvais ; de deux étoiles pour un appariement jugé moyen ; de 3 étoiles pour un appariement jugé bon.

USER2 \ USER1	3*	2*	1*	TOTAL
3*	47	61	9	117
2*	9	36	16	61
1*	0	3	8	11
TOTAL	56	100	33	189

TABLE 5.6 – Comparaison des jugements de pertinence de deux juges

distinctes chez les deux juges.

Les exemples **(9)** et **(10)** correspondent tous deux à des paires automatiques que l'utilisateur 1 a jugées mauvaises, et que l'utilisateur 2 a jugées bonnes. S'agissant du **(9)**, le contenu de la vidéo proposée ne développe pas exactement le sujet de l'article, mais aborde bien le sujet global des primaires Républicaines et ne date que de quatorze jours avant la publication de l'article. Il est difficile de déduire d'un simple jugement binaire ce qui a plu à l'utilisateur 2 dans cette paire et ce qui a parallèlement déplu à l'utilisateur 1, pour que leurs notes respectives soient si opposées. Quant à l'exemple **(10)**, la vidéo relate le fait introduit dans l'article comme la cause du sujet abordé, *i.e.* des semaines de chantier dans le métro. Cette vidéo ne date par ailleurs que de deux jours avant la publication de l'article, ce qui en fait selon nous un bon résultat. Il est encore plus difficile dans ce cas-ci de saisir les raisons de la dépréciation de cet appariement par l'utilisateur 1.

Notons par ailleurs que bien que les utilisateurs aient accès aux contenus exhaustifs des documents lors des évaluations (titre et description), ils se contentent régulièrement de n'observer que leurs titres pour juger de leur similarité. Ce fait pourrait selon nous expliquer certains cas d'écart d'appréciation d'une même vidéo entre le système et un utilisateur. En effet, bien qu'il surpondère les termes du titre, le système exploite l'ensemble du texte d'un document pour sa représentation. Si la description contient des informations essentielles à la compréhension de l'ensemble du document, le système la prendra en compte mais l'utilisateur n'y aura pas accès en se contentant de lire le titre.

(9)
Article (2016-11-21) – *Primaire : NKM, admet qu'elle ne gagnera pas et raille ses rivaux*
Vidéo auto (2016-11-07) – *Primaire de la droite : quand Sarkozy tacle NKM*

(10)
Article (2016-11-21) – *Rennes : Trois semaines d'arrêt pour le chantier du métro suite à un effondrement dans un magasin*
Vidéo auto (2016-11-19) – *VIDÉO - Le plancher d'un magasin s'effondre à Rennes*

Pour quantifier le désaccord entre ces deux juges, nous calculons le coefficient de Kappa

de Cohen (Cohen, 1960). Il mesure la qualité de l'accord réel entre des jugements qualitatifs, par comparaison du taux d'accord observé à la probabilité d'un accord aléatoire. Sur ces données, nous obtenons κ=0.18, soit un taux d'accord très faible au regard du classement proposé par (Landis et al., 1977).

Ce faible accord met en lumière le biais d'une évaluation manuelle, que nous évoquions plus haut, en l'absence de grille d'appréciation commune à tous les juges. Ce constat nous fait relativiser la difficulté de la tâche pour laquelle il faut s'affranchir des critères individuels pour construire un modèle universel, avec le risque de ne pas satisfaire les attentes spécifiques de chacun des utilisateurs. Le fait qu'il s'agisse d'une tâche d'appariement pose dès le début une forte contrainte de défaut d'information en sortie. Par ailleurs, les requêtes soumises au système étant des articles complets, riches d'informations, minimise les chances de proposer un résultat répondant à tous ses aspects parmi un ensemble non exhaustif de vidéos. Mis à part les cas *parfaits* pour lesquels il existe en base une vidéo illustrant exactement l'article, le système doit opérer des choix pour déterminer sur quels aspects de la requête mettre l'accent, puis rechercher une vidéo illustrant ces aspects. Mais si l'utilisateur, à la lecture de l'article, retient des aspects sous-représentés par le système, son besoin risque de n'être que partiellement satisfait, voire pas satisfait du tout.

5.5 Conclusion

Nous sommes revenus dans ce chapitre sur le protocole d'évaluation mis en place chez Mediabong pour évaluer les performances du système développé en réponse à leurs besoins particuliers. Au travers d'exemples issus du corpus, nous avons exposé les contraintes que pose le contexte industriel, relatives à la complexité des exigences et à l'instabilité des jugements de pertinence.

Après avoir démontré la difficulté de mettre en place une évaluation dans un cadre expérimental de RI, nous avons proposé un protocole adapté à un système industriel en conditions réelles d'application. Notre problématique dépassant le cadre de celui d'un SRI *classique*, c'est sur un modèle d'évaluation de classification binaire que s'appuie ce protocole. Il considère la dépendance des résultats à la disponibilité de vidéos en collection, en distinguant les articles susceptibles d'y trouver une vidéo pertinente de ceux pour lesquels trouver une vidéo illustrant leur contenu trop spécialisé est illusoire.

Dans l'intention de comparer les performances de différents systèmes et en l'absence d'ensemble de test, nous avons conçu le protocole en s'inspirant de méthodes d'A/B testing. En tenant compte de la dépendance des performances aux données traitées, nous avons ainsi pu observer et opposer différents systèmes entre eux sur la base de résultats obtenus sur des données comparables.

Malgré les imperfections que ce protocole présente, notamment dû au dynamisme des données qu'il est nécessaire de considérer, il nous a permis de confronter les performances de différents systèmes pour finalement en sélectionner un qui satisfait aujourd'hui pleinement MEDIABONG.

Chapitre 6

Ensemble de test et expérimentations

Parallèlement à l'implémentation du système industriel, nous avons mené une série de tests dans un cadre plus expérimental afin de comparer différentes approches pour la résolution de la tâche d'appariement de contenus qui nous occupe ici.

Pour observer l'impact de certains facteurs sur les performances, nous avons comparé les résultats obtenus par différentes configurations de système. Ces configurations se distinguent selon trois grands types de variables que sont la représentation des documents, le modèle de similarité et la prise en compte ou non du paramètre temporel dans le calcul du score final.

La Figure 6.1 présente les paramètres considérés lors de ces expérimentations, dont l'ensemble des combinaisons permet d'obtenir finalement 36 configurations différentes de système.

Dans le but d'évaluer et de comparer ces configurations, nous proposons un ensemble de test construit en sélectionnant un échantillon représentatif des contenus traités en production. Les données de cet ensemble sont annotées manuellement, et constituent donc une ressource exploitable en l'état pour de futures expérimentations par d'autres chercheurs des communautés de la RI ou du TDT.

Nous décrivons dans ce chapitre la façon dont nous avons fait varier ces différents paramètres, en débutant par les différentes représentations documentaires en Section 6.1 puis les différents modèles de RI en 6.2. Nous revenons ensuite en Section 6.3 sur la fonction implémentée pour saisir la fraîcheur des résultats puis sur la manière de l'intégrer au calcul du score final. La Section 6.4 présente finalement la méthode d'évaluation des performances des différents systèmes, en décrivant les données exploitées et les métriques implémentées puis une discussions sur les résultats obtenus.

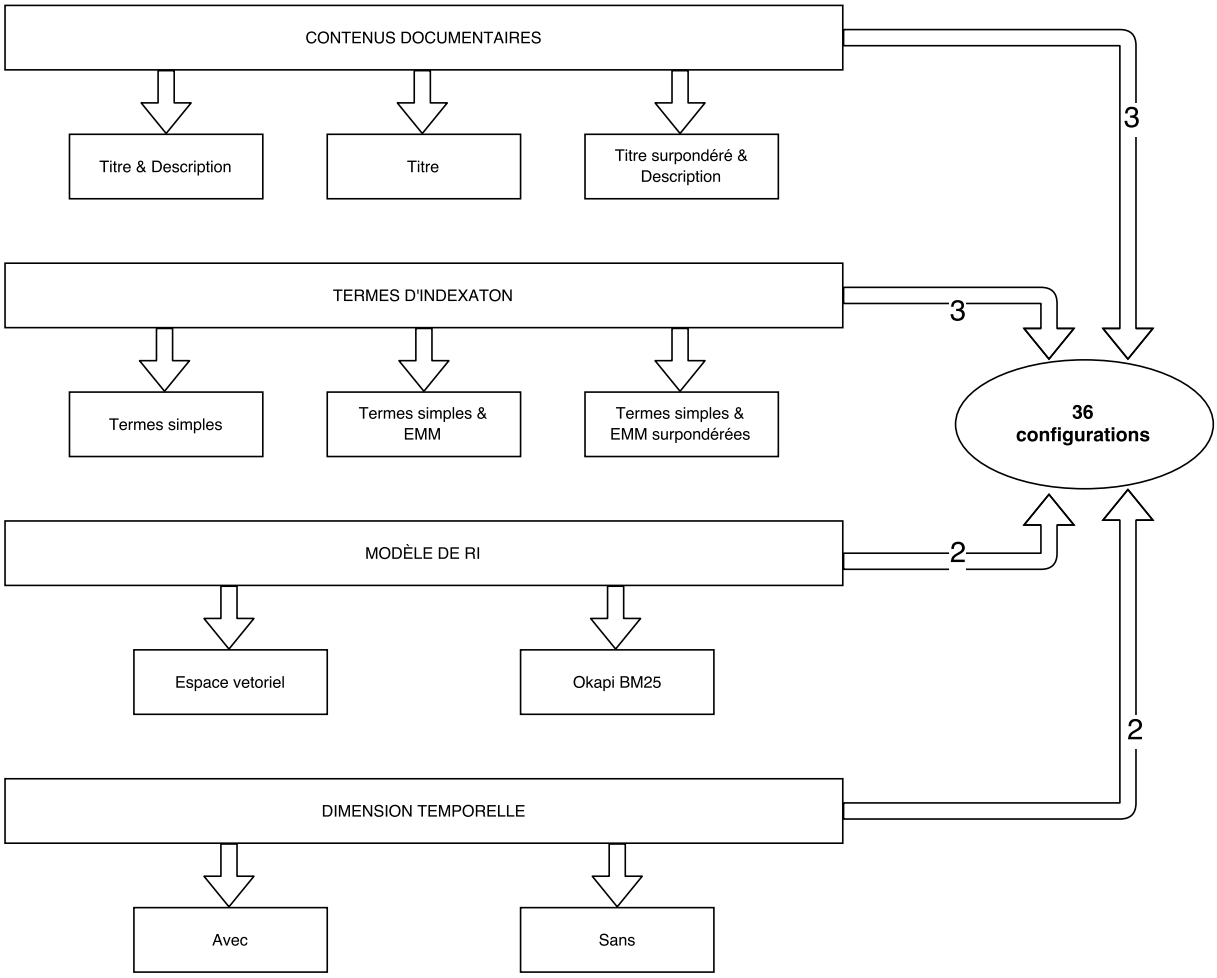

FIGURE 6.1 – Paramètres testés lors des expérimentations

6.1 Représentation documentaire

Nous avons souhaité observé l'influence de la représentation des documents sur les performances du système d'appariement. Ainsi, nous avons distingué deux types de variables : d'une part le segment de texte considéré et de l'autre, le type de termes retenus pour l'indexation.

Les documents traités, articles comme vidéos, sont semi-structurés puisque l'on peut y distinguer le titre de la description. Les titres condensent généralement l'essentiel de l'information décrite dans la suite du document et sont donc susceptibles de contenir des termes particulièrement représentatifs de l'ensemble du contenu. En revanche, leur brièveté fait qu'ils peuvent passer sous silence certaines données potentiellement importantes pour représenter précisément l'information décrite.

En outre, si les titres sont en moyenne tous de même longueur selon les articles (*resp.* les vidéos), les descriptions sont en revanche de tailles très variables, en fonction des diffuseurs (*resp.* producteurs) et de la capacité du *parser* à extraire le contenu des pages

HTML des articles. Le Chapitre 4 présente de façon détaillée et chiffrée ces variations, qu'il est important de souligner puisque la taille des documents considérés est susceptible d'influencer les résultats d'un SRI.

Afin de mesurer l'impact des titres sur les résultats du système, nous confrontons différentes représentations qui se distinguent par le type de section considérée ou le poids accordé à celle-ci :

— Les titres seuls
— Les titres et les descriptions
— Les titres surpondérés et les descriptions

Par ailleurs, les travaux en RI ont largement démontré l'importance de considérer des termes d'indexation plus informatifs que les mots simples seuls. C'est particulièrement le cas lorsque l'on traite des données d'actualité, tel que le démontre les travaux de (Hatzivassiloglou, Gravano et al., 2000) sur une tâche de *clustering* sur le corpus TDT2 [1]. Les auteurs de ces travaux posent l'hypothèse que la description d'événements impliquent l'utilisation de syntagmes nominaux (SN) figés et de nombreux noms propres, permettant à tout lecteur de saisir rapidement de quoi il s'agit. Ainsi, les entités nommées (EN) de type Personne, considérées comme saillantes d'après (Landragin, 2004), devraient avoir plus de poids dans la représentation d'un article à caractère événementiel que les autres termes le composant. L'hypothèse de (Hatzivassiloglou, Gravano et al., 2000) se vérifie dans leurs résultats, puisqu'ils obtiennent de meilleures performances en intégrant et surpondérant ce type de données, même si ce gain reste faible. Cette faiblesse s'explique selon les auteurs par le fait que ces données sont récupérées automatiquement et que les résultats des systèmes en charge de cette tâche peuvent ne pas tout repérer (silence) et/ou extraire des erreurs (bruit).

Sur le même type de tâche, (Moschitti et al., 2004) concluent en revanche que l'ajout d'informations linguistiques à tendance à influer négativement sur les résultats de *clustering*. Ils comparent les performances de systèmes obtenus à partir de différentes représentations documentaires, intégrant notamment des SN et/ou des informations sémantiques extraites de WordNet. Les conclusions démontrent que ni les informations morpho-syntaxiques, ni les informations sémantiques n'aident le système à obtenir de bons résultats. Le risque en considérant des unités poly-lexicales est que même si elles sont plus précises, elles permettent en revanche moins de liberté. L'exemple est pris avec l'unité *George Bush* qui, lorsqu'elle est considérée comme une des dimension du vecteur texte, ne pourra être assimilée à l'unité simple *Bush*, alors qu'une représentation en *sac de mots* aurait permis ce rapprochement.

Souhaitant également évaluer l'impact du type de termes considérés pour l'indexation,

1. *Cf.* Chapitre 2 pour plus de détails sur ce corpus.

nous testons trois représentations différentes des textes :

— Les termes simples
— Les termes simples et les EMM
— Les termes simples et les EMM surpondérées.

Notons que les termes considérés sont à chaque fois récupérés sous leur forme lemmatisée par TreeTagger. Quant aux expressions multi-mots (EMM), il s'agit de SN et d'EN polylexicaux extraits via la chaîne de traitement présentée au Chapitre 4. Ce même chapitre détaille aussi les méthodes de surpondération des termes évoquées ici, à savoir une simple augmentation des valeurs de *TF* dans la fonction de pondération des termes. Ainsi, lorsque nous parlons d'une surpondération des termes du titre, il s'agit de multiplier par 2 la fréquence d'occurrence des termes présents dans les titres ; lorsque nous parlons d'une surpondération des EMM, il s'agit de multiplier par 2 les unités de type Lieu et par 3 celles de type Personne.

6.2 Modèle de RI

Bien que le classique modèle vectoriel de (Salton, 1971) soit régulièrement implémenté dans les systèmes de TDT, nous avons souhaité comparer ses performances à celles d'un autre modèle classique en RI, le modèle Okapi-BM25 (Robertson et Walker, 1994).

Ces deux approches sont décrites en détails au Chapitre 1. Nous noterons simplement que la pondération des termes dans le modèle vectoriel implémenté dans cette série d'expérimentations est une fonction classique *TF*IDF*.

Par ailleurs, nous remarquons que si le score obtenu par le modèle vectoriel est normalisé, puisque le cosinus ne peut varier qu'entre 0 et 1, les scores obtenus par un modèle Okapi-BM25 ne le sont pas. Il s'agit en effet d'une somme de probabilités sur chaque terme de la requête, le score final dépend donc de la taille de la requête soumise en entrée. De ce fait, il est assez compliqué de fixer un seuil de score distinguant les bons résultats des mauvais avec ce modèle, puisqu'un seuil devrait en fait être fixé indépendamment pour chacune des requêtes. Il est d'ailleurs rappelé dans (Piworwarski, 2003) qu'il reste difficile, pour la plupart des systèmes de RI, de fixer automatiquement un seuil discriminant les documents pertinents des documents non-pertinents pour une requête.

Il est en revanche tout à fait possible d'ordonnancer les résultats en fonction des scores obtenus par chacun des modèles et c'est donc sur des métriques intégrant cette notion d'ordonnancement que se baseront nos mesures comparatives. Ce que nous souhaitons comparer dans ces expérimentations est la capacité d'un système à proposer une bonne

vidéo pour un article en tête des résultats. L'objectif à terme est en effet de pouvoir associer automatiquement une unique vidéo aux articles, et il est donc nécessaire d'optimiser la précision au premier document.

6.3 Considérations temporelles

Nous avons évoqué au Chapitre 2 l'importance du paramètre temporel dans le traitement des données d'actualités dans le cadre d'une tâche de TDT. Nous avons par la suite proposé au Chapitre 4 une fonction empirique permettant de saisir cette proximité temporelle entre un article et une vidéo.

Dans le cadre expérimental qui fonde ce chapitre, nous avons opté pour une fonction décrite dans de précédents travaux dédiés spécifiquement à la tache de TDT, (Chy et al., 2015). L'unique variable considérée dans cette fonction, présentée en 6.1, est I : elle correspond à la distance en nombre de jours séparant la date de publication de l'article soumis en requête q à celle de production de la vidéo D_i.

$$score\_date\ (q,D_i) = \frac{1}{\sqrt{(\sqrt{I+1})}} \tag{6.1}$$

Pour mesurer l'impact de ce facteur temporel sur les performances des différents systèmes d'appariement testés, nous l'avons intégré à chacune des 18 configurations déjà existantes. (Chy et al., 2015) proposent de faire la somme de l'ensemble des scores calculés pour chaque paire document-requête comme score final. Dans l'objectif de normaliser ce score final, nous choisissons d'en calculer la moyenne géométrique plutôt que la simple somme. Comme le présente la fonction 6.2, il s'agit de récupérer la racine carré du produit des deux scores calculés pour chaque paire de documents.

$$score\_combined\ (q,D_i) = \sqrt{score\_content * score\_date} \tag{6.2}$$

6.4 Évaluation

6.4.1 Sélection des données

Afin d'obtenir un ensemble de test représentatif de tous les contenus traités, il est recommandé, dans les standards de la RI, de sélectionner aux moins 50 requêtes. Les auteurs de (Voorhees, 2000) démontrent dans leurs travaux que le comportement des mesures est stable à partir de ce nombre de données. Ils précisent en revanche que lorsque

l'on s'intéresse particulièrement aux métriques de haute précision, comme c'est le cas dans nos travaux, ce nombre peut s'avérer insuffisant. Toutefois, dans le cadre industriel imposé à cette thèse, il ne nous a pas été possible de faire évaluer manuellement plus de 50 requêtes, qui représentaient d'ores et déjà pour les juges un investissement en temps conséquent.

Par ailleurs, nous avons affaire à des contenus très hétérogènes qui recouvrent une grande diversité de thématiques. Toutefois, ces thématiques sont inégalement représentées dans les données réelles, notamment parce que l'on traite beaucoup de contenus d'actualité et assez peu de contenus sur des sujets plus spécifiques. Aussi, afin de conserver une certaine représentativité de cette variété dans l'ensemble de test, nous avons sélectionné des articles en fonction de la distribution des thèmes traités en conditions réelles de production.

La Table 6.1 présente cette répartition dans laquelle on distingue en tête de liste (et en gras) les thèmes particulièrement représentés dans nos données et par la suite les thèmes présents mais sous-représentés. On nommera par la suite les premiers thèmes *principaux*, représentant presque 70% de l'ensemble des articles traités tandis que les autres seront dénotés par le terme générique de thèmes *divers*.

INTERNATIONAL	**10**
SOCIETE	**9**
POLITIQUE	**5**
ECONOMIE - FINANCE	**3**
SPORT	**3**
PEOPLE	**2**
SANTE	**2**
ACTUALITE	1
ANIMAUX	1
SCIENCES	1
MAISON	1
MODE	1
INSOLITE	1
ECOLOGIE - ENVIRONNEMENT	1
CUISINE	1
ACTU MEDIA	1
CINEMA	1
MUSIQUE	1
BEAUTE	1
TOURISME VOYAGE	1
ART ET CULTURE	1
HIGH TECH	1
AUTO / MOTO	1

TABLE 6.1 – Distribution des thèmes dans l'ensemble de test

En se basant sur ces critères, nous sélectionnons aléatoirement 50 articles sur un intervalle de trois mois afin de garantir une hétérogénéité des contenus analysés, en évitant les doublons[2]. L'Annexe D présente la liste des titres de ces requêtes.

Par ailleurs, nous avons fixé la collection de vidéos à un instant t, afin de pallier le biais que représente son évolution permanente. Ainsi, au moment de ces expérimentations, c'est une collection de 599 503 vidéos qui est interrogée. En amont de tout traitement par les différents systèmes testés, nous sélectionnons un sous-ensemble de vidéos pour chacune des requêtes sur la base des termes partagés[3]. Nous créons ainsi une sous-collection de vidéos spécifique à chacune des requêtes, comptant en moyenne 87 131 vidéos. Ce sont ensuite ces sous-collections qu'interrogent les différents systèmes testés lors des expérimentations.

6.4.2 Pooling

Chacune des 36 configurations de systèmes traitent chacune des 50 requêtes et propose en sortie l'ensemble des vidéos de la sous-collection liées à l'article, dans l'ordre décroissant de leur score.

Les 50 premières vidéos retournées par chacun des systèmes sont récupérées pour former un ensemble, appelé un *pool*. La taille de ce *pool* peut varier en fonction des requêtes, entre 139 et 340, avec une moyenne de 264. Cette variation s'explique par le fait que pour certaines requêtes, tous les systèmes ont tendance à s'accorder sur les résultats de tête, alors que dans d'autres cas les résultats sont plus hétérogènes selon les configurations.

6.4.3 Jugements de pertinence

Trois juges ont été mobilisés pour la tâche d'évaluation des résultats de l'ensemble de test. De même que pour les autres annotations de ce type dans ces travaux de thèse, les juges sont des salariés de Mediabong, et non des experts dédiés à cette tâche particulière.

Les juges ont accès aux titres des articles et vidéos, ainsi qu'à leurs descriptions. Toutefois, les descriptions pouvant s'avérer longues, elles sont masquées par défaut dans l'interface d'évaluation mais peuvent être consultées au moyen d'un simple clic.

Nous avons cependant constaté que ces descriptions n'étaient que rarement consultées par les juges, qui se contentent la plupart du temps de la lecture du titre pour déduire le sens global du document parcouru. Cette remarque n'est pas anodine, car nous supposons alors qu'une indexation sur les seuls termes du titre, ou leur surpondération, pourrait en

2. Beaucoup de nos partenaires récupèrent des dépêches AFP qu'ils publient parfois en l'état. On peut se retrouver alors à traiter exactement les mêmes contenus provenant de différents partennaires.

3. C'est-à-dire que l'on ne conserve que les vidéos de la collection qui partagent au moins un terme de l'article.

tirer avantage.

La tâche consistait à attribuer un jugement binaire [4] à chacune des vidéos présentées en réponse à un article donné. En moyenne, 264 vidéos sont proposées en réponse à chacun des 50 articles. Ce sont au total 13 127 jugements de pertinence qui ont été attribués par chacun des différents juges engagés. Le temps d'attribution d'un jugement de pertinence à une paire article-vidéo est estimée à environ 3 secondes, élevant le temps nécessaire pour l'évaluation totale à 11 heures de travail. La tâche est répétitive et plus que fastidieuse, raison pour laquelle les juges n'ont accordé en moyenne qu'une dizaine de minutes quotidienne à cette annotation, et ce de façon non assidue. Soumises aux jury au mois de mars 2017, ce n'est donc qu'à la fin du mois de juillet suivant que nous avons pu disposer de l'ensemble de ces données pour les exploiter.

L'annotation se déroule en triple aveugle, c'est-à-dire que chacun des juges évalue chacune des paires article-vidéo composant l'ensemble de test. On garantie ainsi de pallier autant que faire se peut le biais de subjectivité grâce à un système de vote. Le jugement finalement associé à une paire est celui accordé par au moins deux des juges. Par exemple, sur une instance donnée, si deux d'entre eux estiment que l'appariement est bon, on lui attribue finalement l'étiquette *Bon* même si le troisième la juge mauvaise.

Afin de vérifier le taux global d'accord entre les différents juges, nous avions initialement envisagé de calculer le coefficient de Kappa de Cohen (Cohen, 1960). Cependant, la définition de ce coefficient considère la probabilité d'un accord aléatoire comme l'une de ses variables. Dans une situation *classique* d'annotation, les annotateurs disposent d'un protocole strict, sur la base duquel ils élaborent leurs choix d'annotation. Dans ces cas-là, la prise en compte de l'accord aléatoire est pertinente. L'idée sous-jacente est qu'à partir du moment où les juges ne sont pas en accord parfait, c'est qu'une partie de leurs annotations a été produite de façon non maîtrisée [5] car ils ne peuvent être à la fois en désaccord et tous conformes à ce qui est attendu (Mathet et al., 2016). Il est donc important de pouvoir saisir cette proportion d'aléatoire dans les résultats finaux.

Or dans notre cas, aucun protocole n'est fourni aux juges, donc aucune réponse précise n'est attendue pour une instance donnée. Il s'agit de jugements, subjectifs par nature, et l'on souhaite savoir dans quelles proportions les juges sont d'accord entre eux, sans s'intéresser à leur fidélité par rapport à un protocole établi. Aussi, nous préférons calculer l'accord entre les différents juges en observant simplement le nombre de cas où ils sont en accord, par rapport à l'ensemble des cas annotés (*i.e.* l'accord observé).

Les Tables 6.2, 6.3 et 6.4 présentent les matrices confondant les annotations fournies par les juges deux à deux. Les valeurs de Kappa de Cohen (κ) et d'accord observé (α) y

4. Soit *Bon*, soit *Mauvais*.
5. *i.e.* Non maîtrisée par rapport à ce qui été attendu au regard de la norme établie par le protocole.

sont associées à chaque fois. On peut y remarquer que la valeur de K obtenu entre USER1 et USER3 est le plus faible des trois (0.34) alors que leur taux d'accord observé est au contraire le plus haut (0.84). Ceci s'explique par le fait que ces deux juges ont le plus fort taux d'accord sur la classe des résultats *Mauvais*, mais ont parallèlement le plus faible sur la classe *Bon*. Or K pénalise la trop forte concentration des annotations dans l'une des classes, comme c'est le cas ici.

USER2 \ USER1	Bon	Mauvais	TOTAL
Bon	1407	1769	3176
Mauvais	723	9228	9951
TOTAL	2130	10997	13127

K = 0.42, **A = 0.81**

TABLE 6.2 – Accord observé entre USER1 et USER2

USER3 \ USER1	Bon	Mauvais	TOTAL
Bon	812	826	1638
Mauvais	1318	10171	11489
TOTAL	2130	10997	13127

K = 0.34, **A = 0.84**

TABLE 6.3 – Accord observé entre USER1 et USER3

USER3 \ USER2	Bon	Mauvais	TOTAL
Bon	1172	466	1638
Mauvais	2004	9485	11489
TOTAL	3176	9951	13127

K = 0.38, **A = 0.81**

TABLE 6.4 – Accord observé entre USER2 et USER3

Remarquons finalement que les trois juges s'accordent tous entre eux dans 9 575 cas (73%), parmi lesquels 8 918 sont des jugements en *Mauvais* (93%), contre seulement 657 en *Bon* (7%). Les 3 552 restants (27%) correspondent aux cas ou au moins deux des juges étaient en accord, mais pas le troisième. Malgré une certaine diversité inter-personnelle dans ces jugements, ce sont eux qui servent de résultats de référence lors des calculs des différentes métriques rendant compte des performances des systèmes.

6.4.4 Calcul des métriques

Nous nous intéresssons particulièrement dans ces travaux à l'optimisation des précisions hautes, puisque seules les cinq meilleures vidéos retrouvées pour un article sont proposées aux utilisateurs. De plus, l'objectif final est d'associer la première vidéo retournée

à l'article sans validation manuelle. Aussi, nous évaluons et comparons les performances des différentes configurations de systèmes sur la base des mesures de précision à 1 et à 5 documents (soit respectivement *P@1* et *P@5*). Par ailleurs, afin de se faire une idée de leurs performances plus globales, nous choisissons de calculer la métrique de *MAP*, correspondant à la moyenne des précisions moyennes sur l'ensemble des requêtes de test. Le Chapitre 1 expose plus en détails l'utilité et le calcul de chacune de ces métriques.

Nous présentons par la suite les résultats bruts ainsi que des graphiques confrontant les résultats des différentes configurations, pour chacune des métriques :

— *P@1* : Table 6.5 & Graphique 6.2
— *P@5* : Table 6.6 & Graphique 6.3
— *MAP* : Table 6.7 & Graphique 6.4

Sur chacun des graphiques, les courbes en traits pleins correspondent aux configurations intégrant le modèle à espace vectoriel (VSM), tandis que les courbes en pointillés correspondent à celles intégrant le modèle Okapi-BM25. Quant aux couleurs, les courbes vertes illustrent les configurations ne considérant que le titre pour l'indexation ; les bleues le titre et la description ; les jaunes le titre surpondéré et la description. L'axe des abscisses combine les paramètres relatifs au type de termes d'indexation et à la prise en compte de la date dans les scores.

Il est assez difficile de généraliser des conclusions définitives de ces seules valeurs sans considérer les écart-types, puisque nous avons observé que les résultats variaient grandement d'une requête à l'autre : certaines peuvent approcher une haute précision de presque 1, tandis que d'autres tendent au contraire vers 0. Ces derniers cas correspondent aux requêtes pour lesquelles aucun résultat proposé par le système n'a été automatiquement jugé bon par les utilisateurs, ce qui peut arriver dans le cadre de cette étude puisque la collection de documents interrogée n'est pas en mesure de répondre à toutes les requêtes.

Le fait que l'on considère l'ordonnancement des résultats, sans comparer les scores fournis par les systèmes à des seuils, explique qu'on se retrouve avec des résultats assez moyens. Il aurait été intéressant de pouvoir évaluer ces résultats en termes de classifica-

		Titre		Titre + Desc.		**Titre** + Desc.	
		VSM	BM25	VSM	BM25	VSM	BM25
Simple	Date	0.36	0.3	0.34	0.3	0.38	0.26
	No Date	0.54	0.46	0.46	0.5	0.5	0.5
Simple + EMM	Date	0.36	0.32	0.4	0.26	0.44	0.3
	No Date	0.5	0.46	0.44	0.52	0.5	0.48
Simple + **EMM**	Date	0.38	0.3	0.4	0.3	0.44	0.34
	No Date	0.5	0.48	0.5	**0.56**	0.52	**0.56**

TABLE 6.5 – Résultats pour la précision au premier document (*P@1*)

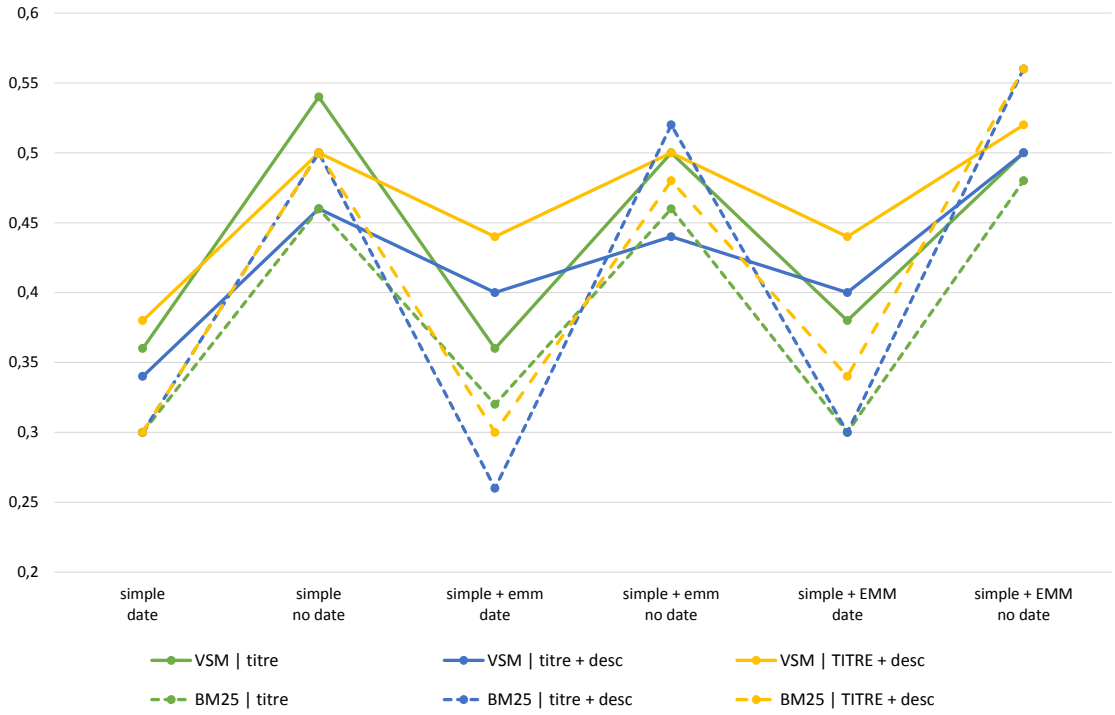

FIGURE 6.2 – Graphique des résultats pour la *P@1*

tion en distinguant strictement les bons des mauvais en réponse à une requête. Mais il nous aurait alors fallu définir un seuil de score pour discriminer ces deux cas, ce qui est difficilement envisageable dans ce contexte expérimental.

Sans généraliser, il est toutefois possible d'observer des tendances nettes à partir de ces résultats quantitatifs. Tout d'abord, deux des systèmes intégrant le modèle Okapi-BM25 présentent la plus haute valeur pour la *P@1* : 0.56. Cette valeur traduit le fait que sur les 50 requêtes de l'ensemble de test, 28 obtiennent un premier résultat pertinent grâce à chacun de ces systèmes. Ils ne sont toutefois pas identiques car en observant le détail des résultats pour chacun des deux, il apparaît que ce ne sont pas les mêmes requêtes qui présentent une *P@1* de 1. La surpondération des termes du titre a donc une influence

		Titre		Titre + Desc.		**Titre** + Desc.	
		VSM	BM25	VSM	BM25	VSM	BM25
Simple	Date	0.352	0.208	0.352	0.208	0.384	0.228
	No Date	**0.492**	0.404	0.472	0.464	0.46	0.472
Simple + EMM	Date	0.368	0.236	0.348	0.208	0.388	0.216
	No Date	0.464	0.412	0.456	0.472	0.448	0.46
Simple + **EMM**	Date	0.384	0.236	0.38	0.22	0.42	0.236
	No Date	0.46	0.424	0.484	0.488	0.48	0.468

TABLE 6.6 – Résultats pour la précision aux 5 premiers documents (*P@5*)

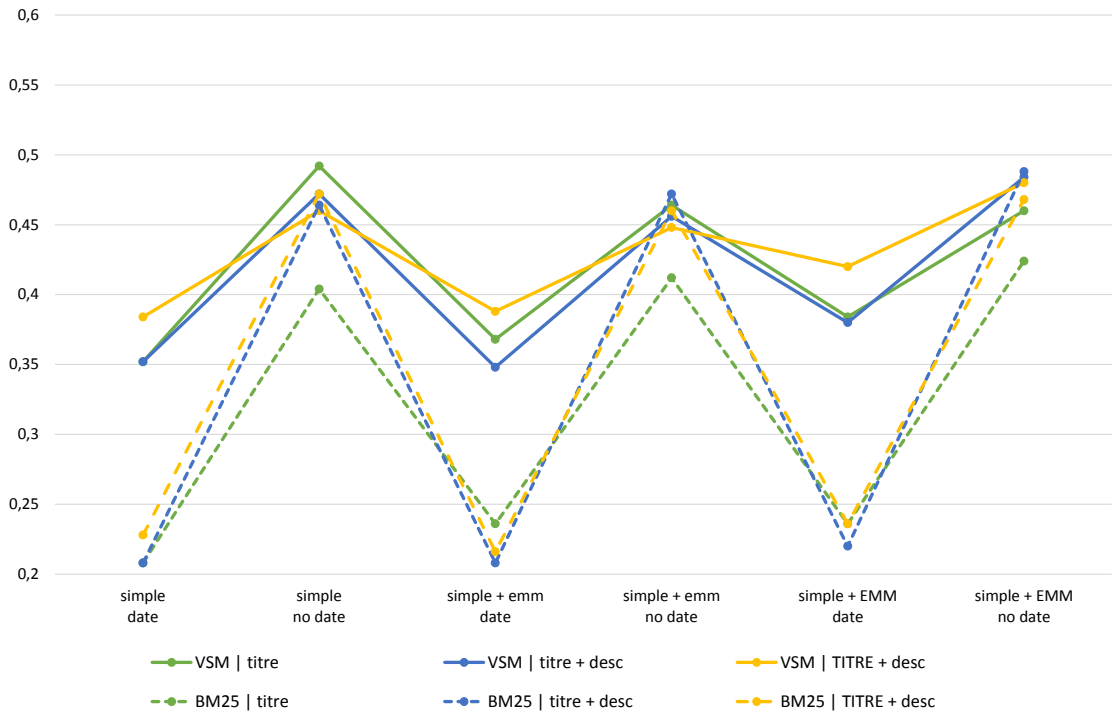

FIGURE 6.3 – Graphique des résultats pour la *P@5*

sur les résultats, mais pas nécessairement positive. En revanche, la prise en compte et la surpondération des EMM a une effet bénéfique sur les performances du modèle Okapi-BM25, et ce pour l'ensemble des trois métriques.

Le système intégrant un modèle vectoriel basé sur les seuls termes du titre présente des performances comparables aux deux précédents, avec une *P@1* de 0.54, rendant compte d'un premier résultat pertinent pour 27 requêtes sur 50. Sur cette même métrique, le modèle vectoriel est par ailleurs meilleur lorsqu'il ne considère que les titres, puisque ses performances chutent lorsqu'il ajoute aux termes de cette section ceux de la description, de 0.04 lorsqu'ils sont surpondérés ($P@1 = 0.5$) et de 0.08 lorsque les termes du titre et de la description sont considérés comme égaux ($P@1 = 0.46$). C'est également la

		Titre		Titre + Desc.		**Titre** + Desc.	
		VSM	BM25	VSM	BM25	VSM	BM25
Simple	Date	0.341	0.251	0.334	0.228	0.354	0.23
	No Date	0.445	0.335	0.419	0.37	0.419	0.393
Simple + EMM	Date	0.345	0.256	0.332	0.212	0.351	0.228
	No Date	0.435	0.339	0.433	0.365	0.436	0.389
Simple + **EMM**	Date	0.367	0.262	0.348	0.242	0.377	0.243
	No Date	0.461	0.375	0.465	0.406	**0.477**	0.436

TABLE 6.7 – Résultats pour la moyenne des précisions moyennes (*MAP*)

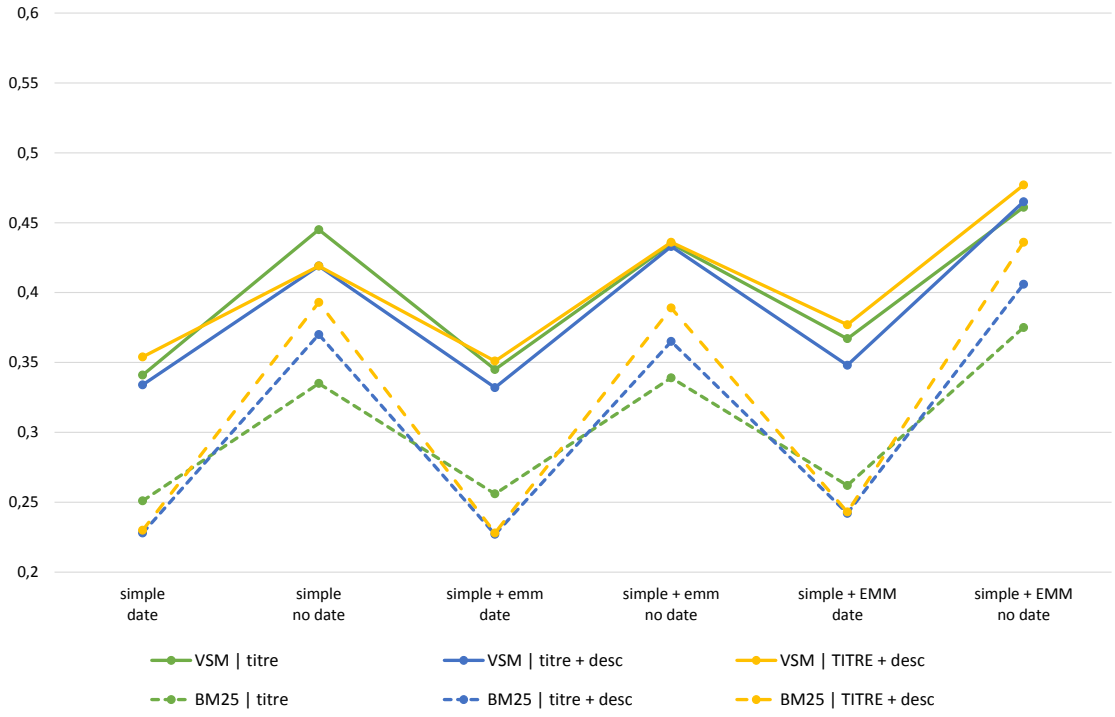

FIGURE 6.4 – Graphique des résultats pour la *MAP*

meilleure des configurations pour la *P@5*, pour laquelle le modèle vectoriel basé sur les titres obtient le plus haut score de 0.492. Toutefois, en prenant en compte les EMM dans les représentations documentaires et en les surpondérant, le modèle vectoriel est meilleur lorsqu'il considère la description et le titre surpondéré, et ce à la fois en termes de haute précision ($P@1 = 0.52$) et de précision globale ($MAP = 0.477$).

Nous remarquons également que l'intégration du paramètre temporel dans le calcul du score final fait systématiquement chuter les performances des systèmes. Il serait cependant hâtif de déduire de cette observation l'inutilité de cette variable dans ce contexte particulier de RI. En effet, les faibles scores affichés pourraient aussi s'expliquer par une mauvaise intégration de ce score dans le score final, plus que par son absence d'intérêt. Par ailleurs, rappelons que ces résultats sont des moyennes sur l'ensemble des requêtes, parmi lesquelles toutes ne sont pas sensibles à la fraîcheur des résultats. Or la majorité des articles traités chez MEDIABONG relèvent de thématiques d'actualité sensibles à ce paramètre.

Ce dernier constat nous a incité à procéder à une évaluation distincte en fonction des thèmes des articles. Nous souhaitions savoir si les résultats étaient comparables selon que le thème de l'article correspond à un thème *principal* ou a un thème *divers*[6]. Nous supposions que les requêtes relatives aux thèmes *principaux* devaient obtenir de meilleurs résultats,

6. *cf.* Table 6.1 qui présente la répartition des thèmes dans ces deux catégories.

s'agissant régulièrement d'actualités pour lesquelles la probabilité de trouver au moins une vidéo pertinente dans la collection est forte. À l'inverse, les requêtes associées aux thèmes *divers* sont généralement atemporelles et/ou spécifiques, et ont moins de chances de trouver un résultat satisfaisant en collection. Or dans l'actuel contexte d'évaluation, les requêtes avec peu – voire pas – de résultats de référence pertinents[7] sont l'un des principaux facteurs des faibles performances des systèmes.

Afin d'observer l'éventuelle influence de ce paramètre temporel sur les résultats des systèmes, nous avons donc calculé les mêmes métriques en distinguant deux ensembles de données : le premier contient les 34 requêtes associées aux thèmes *principaux* ; le second aux 16 requêtes associées aux thèmes *divers*. Les trois pages suivantes présentent les résultats de chacun de ces sous-ensembles de test obtenus pour chacune des trois métriques : Graphique 6.5 pour la *P@1* ; Graphique 6.6 pour la *P@5* ; Graphique 6.7 pour la *MAP*.

On observe de façon flagrante que les tendances se distinguent en fonction du sous-ensemble considéré. Concernant la *P@1*, les résultats de l'ensemble de requêtes liées aux thèmes *principaux* varient entre 0.38 et 0.62, tandis que les valeurs obtenues par le second ensemble de données n'excèdent pas 0.5. Les systèmes à modèle Okapi-BM25 intégrant les termes de toutes les sections (titre et description) et de tous les types (simples+EMM) présentent même une valeur nulle pour cette métrique. Cela signifie que sur l'ensemble des 16 requêtes composant cet ensemble, aucune ne présente un premier résultat pertinent lorsqu'elles sont traitées par ces systèmes. Cette tendance s'observe également pour les résultats de la *P@5* pour lesquels ces mêmes systèmes ne dépassent pas la valeur de 0.05.

Nous observons parallèlement, pour les requêtes des thèmes *principaux*, que le facteur temporel n'est pas aussi néfaste que ce que les premières observations nous amenaient à penser. Le système à modèle vectoriel intégrant tous les termes, surpondérant ceux du titre et ceux de type EMM, présente une valeur *P@1* de 0.559 sans considérer la date, contre une valeur de 0.588 lorsque ce paramètre est intégré au calcul du score des vidéos.

Les résultats de ces analyses comparatives tendent donc à montrer qu'il existe bien une différence entre les articles en fonction des thématiques dont ils relèvent. De ce fait, il serait pertinent de considérer ce paramètre dans le SRI industriel, en envisageant par exemple différents parcours de RI selon le thème de l'article soumis en requête. Une alternative pourrait être envisagée au niveau de l'apprentissage du classifieur distinguant les bons résultats des mauvais : il serait intéressant d'observer les résultats de classification intégrant la thématique comme trait définissant les instances et de les comparer à ceux du classifieur intégré au système actuellement en production.

7. Celles pour lesquelles toutes les vidéos proposées, ou presque, ont été jugées mauvaises par les évaluateurs.

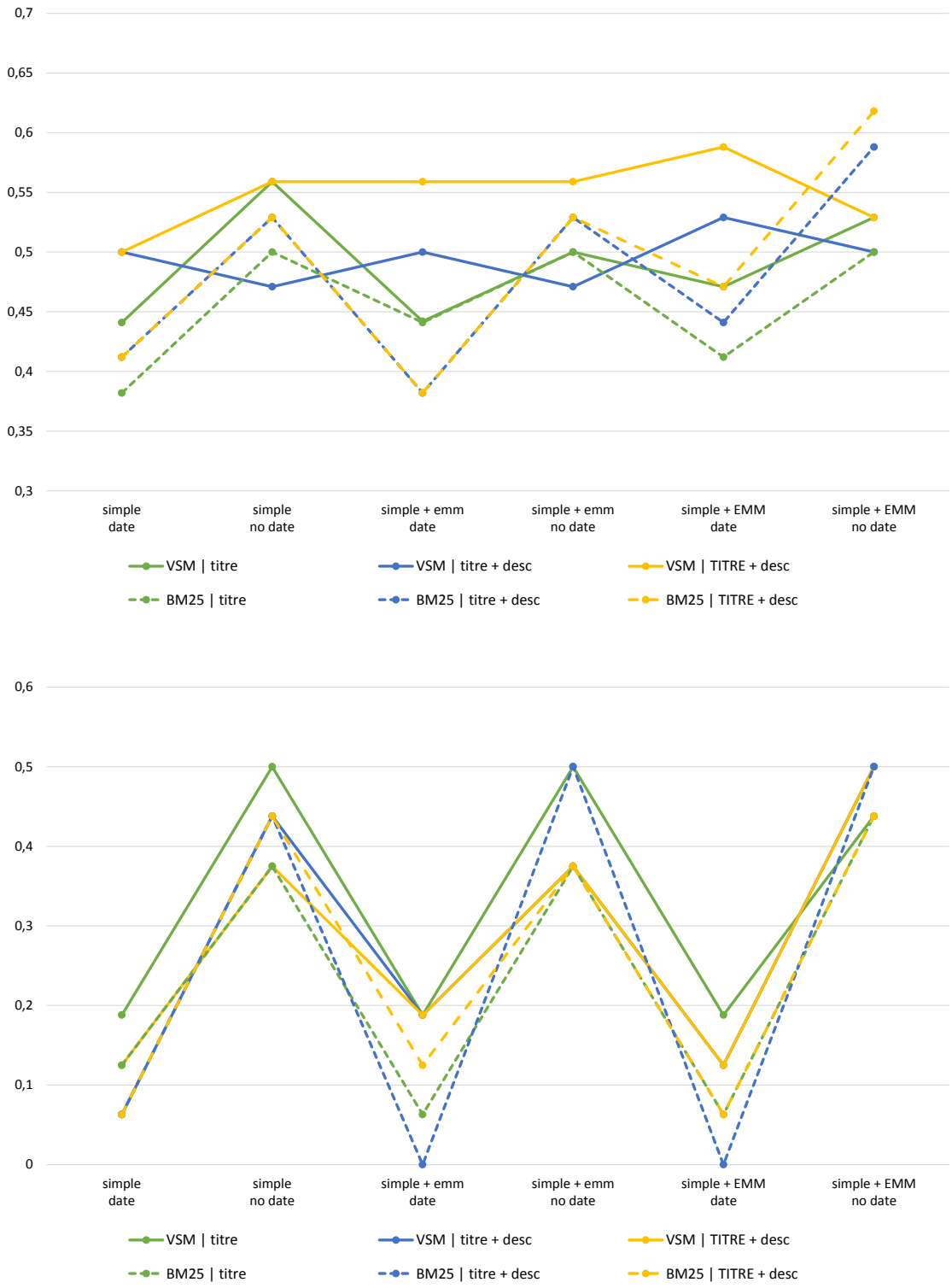

FIGURE 6.5 – Graphiques des résultats de la *P@1* pour les requêtes relevant des **thèmes principaux** (en haut) et des **thèmes divers** (en bas)

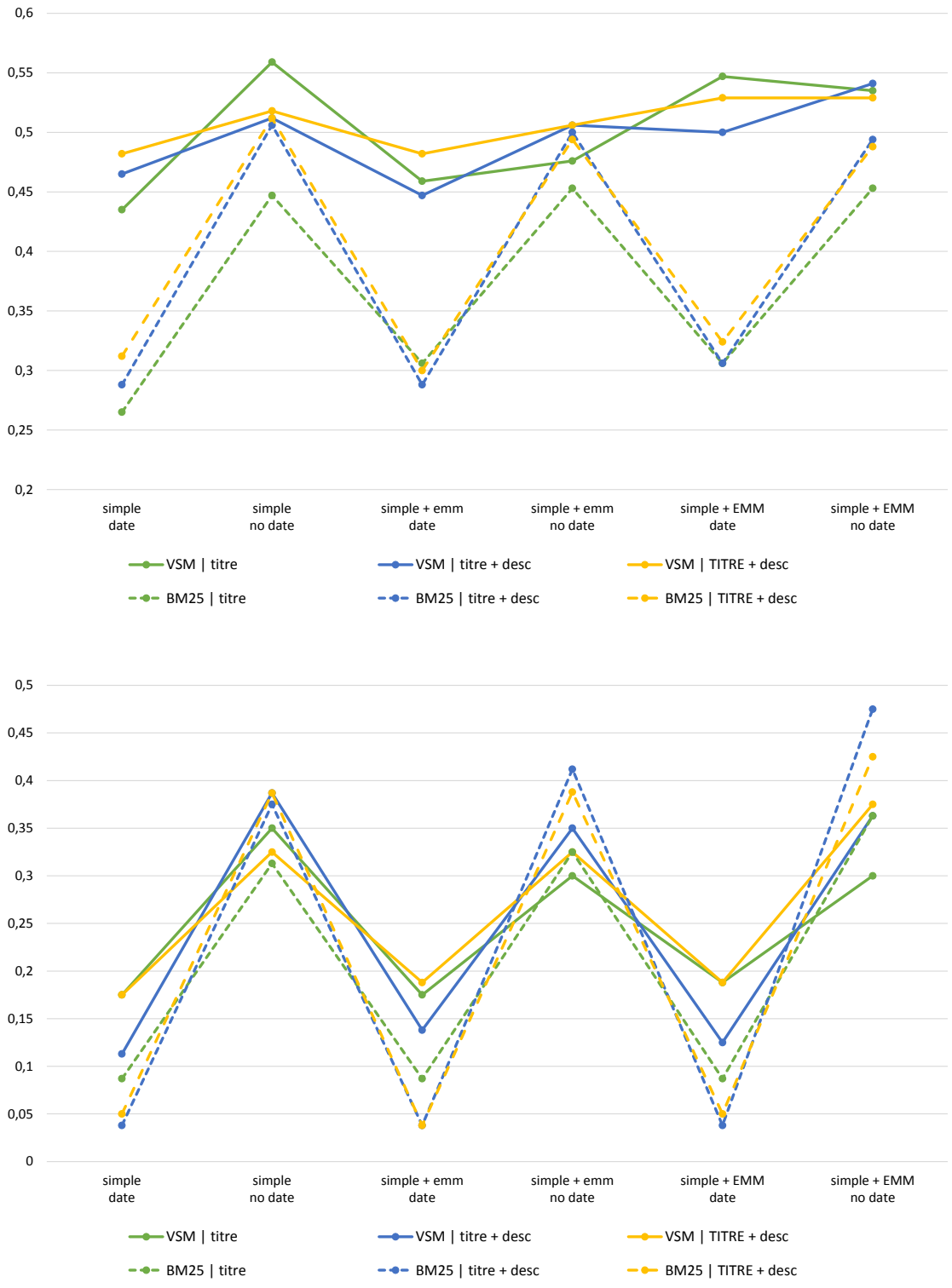

FIGURE 6.6 – Graphiques des résultats de la *P@5* pour les requêtes relevant des **thèmes principaux** (en haut) et des **thèmes divers** (en bas)

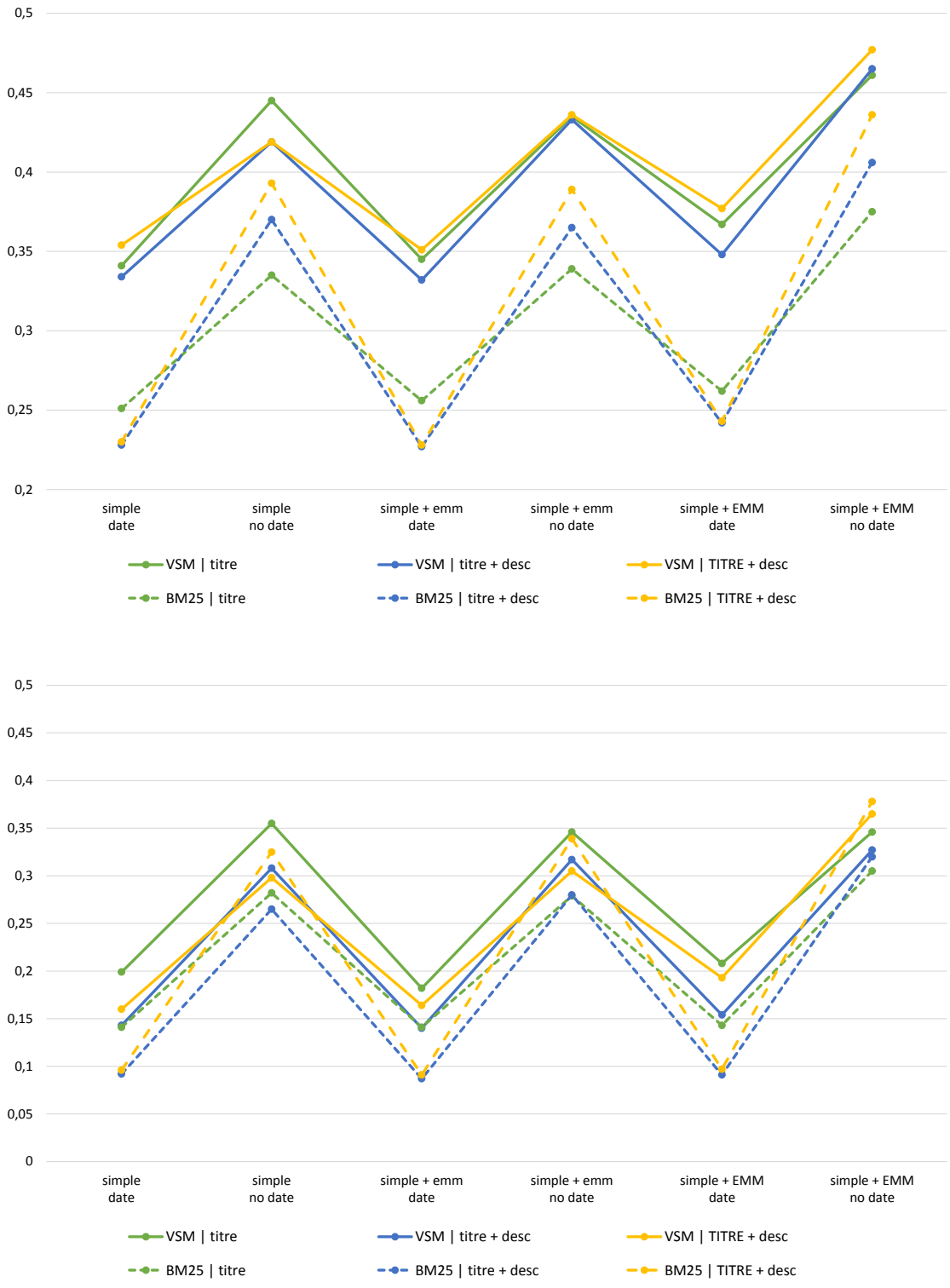

FIGURE 6.7 – Graphiques des résultats de la *MAP* pour les requêtes relevant des **thèmes principaux** (en haut) et des **thèmes divers** (en bas)

6.5 Conclusion

Ce chapitre a présenté la série d'expérimentations que nous avons menée en parallèle du développement du système industriel. Nous souhaitions observer l'influence de différents paramètres sur les résultats de RI, et plus spécifiquement sur ceux de la tâche particulière d'appariement d'articles et de vidéos.

Nous avons construit de bout en bout un ensemble de test suivant les normes des évaluations classiques en RI. C'est donc un ensemble de 50 articles, correspondant aux requêtes, qui est utilisé pour tester ces différentes configurations. Les résultats fournis par chacune d'elles sont filtrés pour former un *pool* évalué manuellement par différents juges. Cet ensemble de test a vocation à être mis à disposition des membres de la communauté RI pour de nouvelles expérimentations.

À partir de ces données, nous avons comparé les performances de 36 configurations de systèmes, se distinguant selon trois types de paramètres : la représentation des contenus, le modèle de RI implémenté et la considération de la composante temporelle dans le calcul du score final. Nous avons ainsi pu conclure que les modèles surpondérant les termes du titre par rapport à ceux de la description, voire ne considérant qu'eux seuls, obtiennent globalement les meilleurs résultats. Par ailleurs, le modèle vectoriel s'est révélé plus performant que le modèle Okapi-BM25, bien que celui-ci considère la taille des documents analysés, plus que variable dans nos données.

Il s'est finalement avéré que la considération temporelle tendait à faire baisser les performances de tous les systèmes. Toutefois, cette conclusion a été infirmée en procédant à une évaluation distinguant les requêtes selon les thématiques auxquelles elles sont associées. Les articles aux thématiques sur-représentées dans le flux de données traité en production décrivent régulièrement des contenus d'actualité pour lesquels il est primordial d'associer une vidéo récente. En ne considérant que ce type d'articles lors de l'évaluation, il est ressorti que la configuration présentant les meilleures performances intégrait le paramètre temporel.

Les paramètres optimaux révélés dans ces expérimentations sont ceux intégrés dans le système en production, Semiabong. Tel que décrit au Chapitre 4, ce système se base sur un modèle vectoriel et offre une surpondération à certains termes d'index en fonction de leur type (SN et EN polylexicales *vs.* termes monolexicaux) et de leur position dans le document (titre *vs.* description). Par ailleurs, bien qu'elle ne soit pas bénéfique à toutes les requêtes, la composante temporelle y est intégrée en fin de chaîne de traitement afin de favoriser les résultats récents. Ceux-ci sont en effet nécessaires en réponse aux articles d'actualité qui forment la majorité des contenus traités.

Conclusion et perspectives

Nous avons abordé dans cette thèse la problématique de l'appariement de contenus textuels médiatiques ou plutôt, l'automatisation de ce processus complexe. S'inscrivant dans un contexte à la fois académique et industriel, nous nous sommes inspirés de l'état de l'art de différents domaines de recherche pour construire un système destiné à l'appariement d'articles en ligne et de vidéos d'information décrites textuellement.

Les conclusions présentées ici reprennent et synthétisent les différents chapitres composant ce mémoire de thèse. Nous proposons en premier lieu un retour sur la discussion liminaire abordée en Partie I. Nous poursuivons par une première synthèse de l'état de l'art exposé en Partie II, puis une seconde concernant nos contributions développées en Partie III. Finalement, nous élargissons ces conclusions en suggérant quelques perspectives de recherche que ces trois ans de travaux ont contribuer à ouvrir.

Retour sur la discussion liminaire

Nous avons introduit ce mémoire par une discussion sur la notion d'*appariement* qui fonde ce travail de recherche. Les travaux menés au cours de cette thèse nous ont en effet interrogés sur cette notion plus complexe que nous ne l'avions initialement envisagé.

Ainsi, nous sommes revenus sur la définition du principe de *similarité* qui sous-tend l'opération d'appariement. Estimer la similarité entre différents objets est une tâche complexe puisque les objets que l'on cherche à comparer sont régulièrement multidimensionnels. C'est notamment le cas des objets textuels, dont la structure repose sur des composants linguistiques de différents niveaux. Nous nous sommes alors demandé si certains de ces composants n'étaient pas privilégiés dans la représentation des textes pour estimer leur degré de similarité, autrement dit, si certains indices linguistiques n'étaient pas saillants dans une telle opération. L'observation de données de notre corpus a démontré l'importance particulière accordée par les juges humains à la dimension sémantique des textes.

Nous avons poursuivi cette discussion autour de la notion de *pertinence*, régulièrement convoquée pour apprécier un résultat fourni par un SRI en réponse à une requête. L'état de l'art en sciences de l'information, en proposant plusieurs définitions, a démontré la difficulté de définir intrinsèquement la pertinence, hors de tout cadre applicatif. Il est apparu

en effet que la pertinence reposait sur un ensemble de critères et qu'il s'avérait difficile pour un résultat automatique de les satisfaire tous. Toutefois, la similarité thématique – *i.e.* le fait qu'un document aborde le même sujet que la requête soumise au système – s'est révélée être un facteur prépondérant pour la pertinence.

Synthèse de l'état de l'art

Nous nous sommes attachés dans la seconde partie de ce mémoire à présenter l'état de l'art des différents domaines relatifs à nos problématiques de recherche.

Ayant considéré la similarité thématique comme l'un des composants fondamentaux de l'appariement article-vidéo, le premier domaine que nous avons exploré est celui de la RI. Ce domaine cherche en effet à répondre au besoin d'information formulé par un utilisateur sous forme de requête en lui proposant des documents estimés pertinents. L'automatisation d'un tel processus implique différentes étapes que sont l'indexation des documents de la collection considérée, la représentation de la requête puis la recherche de documents similaires à celle-ci. Nous sommes alors revenus sur les propositions de différents travaux dans le domaine pour répondre aux questions soulevées à chacune de ces étapes de développement de SRI.

Il est toutefois apparu que la similarité thématique ne pouvait être l'unique critère considéré pour traiter des contenus d'actualité et que la dimension temporelle était également primordiale pour opérer une comparaison de ce type de données. Aussi, le second domaine de recherche que nous avons étudié est celui du *Topic Detection and Tracking* qui s'intéresse à ces textes particuliers marqués temporellement. Parmi l'ensemble des tâches relatives à ce domaine, nous nous sommes particulièrement intéressés à celle de *Story Link Detection* qui cherche à saisir la similarité entre deux nouvelles d'actualité en combinant des méthodes de RI et des mesures de proximité temporelle.

Nous nous sommes finalement intéressés aux systèmes de recommandation qui cherchent à modéliser les préférences de leurs utilisateurs, en fonction de ce qu'ils connaissent d'eux, afin de satisfaire au mieux leurs besoins. L'idée est de construire un SRI personnalisé pour chaque utilisateur grâce aux retours de pertinence que ceux-ci font des items présentés par le système. Les méthodes proposées dans l'état de l'art pour résoudre ce type de tâche nous ont intéressées car elles permettent d'optimiser la combinaison de différentes caractéristiques définissant les items à recommander pour s'approcher aux mieux des préférences des utilisateurs. Or dans notre contexte, nous souhaitions trouver le meilleur compromis entre les mesures de similarité thématique et de proximité temporelle décrivant les paires article-vidéo, en se basant sur les jugements des utilisateurs face à des paires produites automatiquement.

Les conclusions de cet état de l'art nous ont permis de définir plus formellement notre problématique d'appariement de contenus textuels médiatiques. Au regard de ce qu'il nous a appris, nous avons pu opérer certains choix quant au développement du système de résolution de cette tâche, dont les apports principaux sont résumés par la suite.

Synthèse des contributions

La contribution générale de cette thèse est le développement d'un système d'appariement d'articles et de vidéos répondant aux besoins de l'entreprise partenaire, Mediabong. Après trois années d'expérimentations, le système actuellement en production, baptisé Semiabong, permet d'associer automatiquement une vidéo à un article dans 72% des cas en moyenne, ce qui satisfait Mediabong dont l'objectif fixé pour cette tâche était de 70%.

À chaque étape du développement, l'état de l'art ainsi que nos propres analyses ont orienté les décisions prises pour optimiser le système, en tenant compte des contraintes divergentes du besoin complexe de Mediabong.

Représentation documentaire

L'indexation des documents, en l'occurrence articles et vidéos, est l'étape préalable et primordiale d'un SRI. Elle permet de transformer les textes, non-structurés, en une représentation structurée, généralement une liste de mots-clés qu'il convient de sélectionner correctement si l'on souhaite conserver le sens initial du texte indexé.

Afin de dépasser la représentation classique en *sac de mots*, nous avons développé un algorithme d'extraction d'expressions multi-mots (EMM) dont les sorties sont soumises à une vérification manuelle avant d'intégrer une ressource terminologique propriétaire de Mediabong. Cette ressource contient essentiellement des EMM référentielles, telles que des entités nommées (EN) de type Personne et Lieu ainsi que des syntagmes nominaux (SN) regroupés sous le type Divers. Lors de l'indexation d'un document, cette ressource est interrogée pour récupérer les termes qui y figurent. Ces termes particuliers sont surpondérés dans les représentations documentaires, par rapport aux noms, verbes, adjectifs et abréviations monolexicaux récupérés après segmentation du texte.

Par ailleurs, les documents traités dans ces travaux sont semi-structurés puisque l'on peut y distinguer entre autres le titre de la description, *i.e.* le corps de texte plus ou moins long présentant le sujet du document. Nous avons également exploité ces informations de structure pour accorder un poids supplémentaire aux termes du titre par rapport à ceux de la description.

Mesures de similarité

Nous avons choisi de reconsidérer notre problématique d'appariement article-vidéo comme une tâche de RI où l'article constitue la requête fournie au système et la vidéo appariée le meilleur résultat retourné par le système pour cet article.

L'implémentation d'un modèle à espace vectoriel (*VSM*) nous a permis de saisir le degré de similarité thématique entre un article et une vidéo. Cette similarité a en effet été considérée comme l'un des deux facteurs du calcul de similarité globale entre des contenus médiatiques majoritairement ancrés dans l'actualité. Ainsi, pour chaque vidéo candidate pour l'appariement à un article, on calcule ce que l'on note le *score_thematique*.

Le second facteur impliqué relève de la dimension temporelle des contenus puisque nous avons observé que deux documents relatant des sujets proches n'étaient pas nécessairement associables, en tout cas pas pour Mediabong qui a particulièrement insisté sur le besoin de fraîcheur des résultats retourné par le système. En réponse à cette demande, nous avons considéré la dimension temporelle comme une variable participant au contexte de la requête et proposé une fonction empirique pour mesurer le degré de proximité temporelle entre un article et une vidéo. Considérant comme seule variable la distance en nombre de jours entre un article et une vidéo, cette fonction logarithmique accorde un haut score aux vidéos datant du jour ou de la veille de la publication de l'article sans toutefois négliger les plus anciennes en ne leur attribuant qu'un très faible score. Le résultat de cette fonction pour chaque vidéo candidate constitue son *score_date* relativement à l'article.

Nous avons finalement cherché à optimiser la prise en compte de ces *score_thematique* et *score_date* dans un score final reflétant la similarité globale entre un article et une vidéo. Pour se faire, nous avons recueilli des paires article-vidéo automatiquement générées par le système que nous avons soumises à une évaluation manuelle binaire. Nous disposions alors d'un ensemble de paires décrites par leurs deux scores et étiquetées par le jugement de pertinence accordé par l'utilisateur, *i.e.* soit *Bon* soit *Mauvais*. Ces données ont été exploitées comme instances d'apprentissage d'un *SVM* dont le modèle résultant permet d'ordonnancer les vidéos candidates pour un article et d'exclure celles dont le score final est trop faible.

Évaluation des performances

Afin d'évaluer les performances de Semiabong en conditions réelles d'utilisation, nous avons proposé un protocole considérant l'ensemble des contraintes posées par les données traitées et la tâche particulière à accomplir.

Nous avons commencé par exposer les difficultés que posait cette étape d'évaluation, à cause desquelles il nous a été impossible d'utiliser le paradigme de Cranfield qui reproduit

en conditions expérimentales un processus de RI. La principale de ces difficultés est que le besoin d'information n'est pas l'unique besoin à satisfaire dans notre cadre industriel, qui nécessite également de maximiser le nombre d'intégration vidéo afin d'augmenter le revenu généré. Dans cette perspective, un bon système automatique n'est pas uniquement un système capable d'apparier une vidéo pertinente aux articles, mais aussi un système capable d'apparier une vidéo dans un maximum de cas, même si elle s'avère peu pertinente.

Le protocole que nous avons proposé ne nécessite donc pas d'ensemble de test statique servant de résultats de référence auxquels confronter ceux du système à évaluer. Le fait qu'un article n'ait pas nécessairement de résultats pertinents dans la collection de vidéos nous a fait reconsidérer l'évaluation d'un point de vue classificatoire où l'on a distingué binairement l'ensemble des articles pour lesquels il existe au moins une vidéo pertinente en base de ceux pour lesquels aucune n'existe. En comparant les sorties du système aux choix *a posteriori* opérés manuellement par les utilisateurs sur un ensemble d'articles, nous avons recouru à des métriques reflétant les performances globales du système. Selon ces métriques, nous avons pu comparer différents systèmes et sélectionner le plus optimal qui tourne en production depuis fin 2016 et remplit les besoins exprimés par Mediabong en début de projet.

Dans un cadre plus expérimental, nous avons souhaité évaluer et comparer les résultats de différents systèmes du seul point de vue de leur capacité à répondre au besoin d'information. En faisant varier différents paramètres relevant de la représentation documentaire, du modèle de RI et de la prise en compte de la dimension temporelle dans le score d'appariement, nous avons tester 36 configurations différentes en comparant leurs performances respectives obtenues sur un ensemble de test construit à cet effet. Cet ensemble, conforme au standards de la RI, est voué à être mis à disposition de la communauté scientifique pour de futures expérimentations.

Perspectives de recherche

Malgré les performances satisfaisantes qu'offrent Semiabong, le système reste perfectible et toutes les pistes pour automatiser l'appariement article-vidéo sont loin d'avoir été toutes explorées. Les propositions présentées ci-dessous constituent des perspectives que nous supposons capables d'améliorer le système existant, à différentes étapes de son processus.

Au niveau de l'indexation, les choix opérés dans le système actuel relèvent plus de la fouille de textes que du Traitement Automatique des Langues (TAL) puisque nous avons utilisé les informations déduites de la masse de données textuelles dont nous disposions plutôt que les résultats d'analyses ou sorties d'outils linguistiques. Afin de poursuivre

la discussion de l'apport de la linguistique au TAL et plus encore celle de l'apport du TAL à la RI dans lequel nous nous positionnons, nous pourrions envisager d'intégrer au système des informations linguistiques plus fines et confronter ses résultats à ceux du système actuellement en production. La discussion liminaire a en effet mis en lumière un certain nombre d'indices de saillance linguistique, que nous n'avons pas tous exploités dans SEMIABONG. Nous pensons par exemple qu'une analyse syntaxique pourrait aider à améliorer le système en détectant le sujet des phrases, considérés comme saillants par rapport aux objets, puis en leur accordant un poids supplémentaire dans les représentations vectorielles. D'un point de vue sémantique, l'appel à des ontologies ou des outils d'*entity linking* pourrait affiner la ressource terminologique de MEDIABONG en liant certaines entrées. Il serait alors possible d'associer des contenus au sujet des mêmes entités même si celles-ci ne sont pas dénotées par les mêmes unités dans les textes (***e.g.*** *Margaret Thatcher* | *La Dame de fer* ; *Centre Georges-Pompidou* | *Beaubourg*).

En l'état actuel des choses, la mise à jour de la ressource terminologique appelée lors de l'indexation est semi-automatique, puisqu'elle nécessite une annotation manuelle des candidats proposés par l'algorithme d'extraction d'EMM. Disposant, grâce à ces annotations, d'un ensemble conséquent de candidats pour lesquels des juges humains ont (1) décidé s'ils méritaient d'intégrer la ressource et (2) associer un type parmi PERSONNE, LIEU ou DIVERS, nous pourrions envisager d'apprendre un modèle de classification permettent d'automatiser complètement le remplissage de la ressource. Plus précisément, les termes candidats constitueraient les instances d'apprentissage du modèle, dont la classe associée serait NON TERME si l'utilisateur l'a rejeté ou, dans le cas contraire, le type qui lui a été associé (PERSONNE, LIEU, DIVERS). Les traits décrivant ces instances pourraient quant à eux intégrer la fréquence d'occurrence du terme candidat en corpus, la présence de majuscules, le contexte immédiat (*i.e.* termes précédents et suivants) ou encore la présence d'un des constituants dans une liste de noms propres.

Concernant finalement l'évaluation des performances d'un point de vue industriel, nous avons travaillé sur les jugements d'une équipe de salariés de MEDIABONG qui, s'ils ont le mérite d'exister, ne sont pas révélateurs de l'appréciation réelle des résultats par les internautes des sites de presse partenaires. Ce que nous pourrions envisager, à l'instar de l'état de l'art des systèmes de recommandation, c'est la mise en place d'un système de récupération des données des utilisateurs réels du système, c'est-à-dire les lecteurs des articles dans lesquels les vidéos sont intégrées. En comptant le nombre de fois où la vidéo a été visionnée, par rapport au nombre total de visites sur la page, nous pourrions déduire dans quelle mesure la vidéo appariée a été appréciée par les utilisateurs. Ceci ne constituerait que le critère élémentaire pour déduire le niveau d'appréciation de la vidéo mais des indices plus précis, implicites et explicites, pourraient également être envisagés. Parmi les indices implicites, on pourrait par exemple récupérer les informations concernant

le temps de lecture de la vidéo, la position du lecteur dans la page ou encore le fait qu'un utilisateur partage la page ou non après avoir vu la vidéo. Afin de fiabiliser les résultats récupérés, il serait encore davantage intéressant d'avoir accès à des jugements explicites, et l'on pourrait en ce sens envisager de proposer aux internautes une échelle de notation dont le niveau choisi répondrait à une question du type *« Dans quelle mesure avez-vous apprécié cette vidéo ? »*.

Annexes

Annexe A

Articles : Exemples du corpus

Cette annexe présente 10 exemples d'articles traités par Semiabong, reflétant la variété de leurs contenus :

Article 1

Titre	Le Conseil constitutionnel valide le CETA
Chapô	Le Conseil constitutionnel a estimé lundi que l'accord économique et commercial entre l'Union européenne et le Canada (CETA) était compatible avec la Constitution française.
Description	A la suite d'une saisine par plus de 60 députés, le Conseil constitutionnel a estimé, dans une décision datant du 31 juillet, que l'accord de libre-échange entre l'Union européenne et le Canada (CETA) ne nécessitait "pas de révision de la Constitution". Le texte avait été adopté le 15 février par le Parlement européen. Il vise à créer un marché élargi pour les marchandises et les services entre l'UE et le Canada. Il doit désormais être ratifié par les différents pays. Seulement, cet accord ouvre la voie à l'importation de produits canadiens, avec la suppression plus de 99% des droits de douane avec Ottawa. Aussi, les députés derrière cette saisine soulevaient en effet quatre motifs : les "conditions essentielles d'exercice de la souveraineté nationale", le "principe d'indépendance et d'impartialité des juges", le "principe d'égalité devant la loi" et "le non-respect du principe de précaution". En France, le gouvernement a nommé début juillet un groupe d'experts chargé de remettre un rapport début septembre, destiné à évaluer l'impact de ce traité de libre-échange sur l'environnement et la santé.
Tags	∅
Date	2017-08-03
URL	https://www.jolpress.com/le-conseil-constitutionnel-valide-le-ceta-article-836741.html

Éditeur	JolPress

Article 2

Titre	Marseille : une voiture fonce sur des abribus
Chapô	Le conducteur a été arrêté sur le Vieux-Port. Selon le procureur de Marseille, les policiers s'orientent sur "la piste psychiatrique."
Description	Une voiture a foncé dans deux abribus, ce lundi main, à Marseille. Une personne est morte, renversée, une autre a été blessée. Le conducteur, âgé de 35 ans, a été arrêté quelques instants plus tard sur le Vieux-Port.
Tags	∅
Date	2017-08-21
URL	http://www.atlantico.fr/pepites/marseille-voiture-fonce-abribus-mort-3142209.html
Éditeur	Atlantico

Article 3

Titre	Sondage : Allez vous changer votre consommation d'oeufs suite au scandale du Fipronil ?
Chapô	∅
Description	Depuis 2 semaines, le scandale sanitaire des oeufs contaminés au fipronil s'étend et touche aujourd'hui 16 pays européens. Insecticide toxique, le Fipronil peut, à forte dose, endommager les reins, le foie et la thyroïde. En France, 250.000 oeufs bio ou non, seraient concernés et commercialisés depuis avril. Petits producteurs, poules dans votre jardin, régime vegan... : quelles sont vos solutions face au risque sanitaire ? Dites-nous tout dans les commentaires.
Tags	∅
Date	2017-08-15
URL	https://www.consoglobe.com/sondage-consommation-oeufs-fipronil-cg
Éditeur	ConsoGlobe

Article 4

Titre	Faut-il rendre le pourboire obligatoire ?
Chapô	∅
Description	Inquiète de la baisse des pourboires dans les cafés et restaurants, l'Union des métiers et des industries de l'hôtellerie (UMIH) réfléchit à les rendre obligatoires comme cela se pratique dans certains pays. Ses représentants comptent faire une proposition aux députés en ce sens. Légitime ? À vous de juger.
Tags	∅
Date	2017-08-17
URL	http://rmc.bfmtv.com/emission/faut-il-rendre-le-pourboire-obligatoire-646843.html
Éditeur	RMC-BFMTV

Article 5

Titre	Etats-Unis : Les Californiens vont bientôt pouvoir trinquer chez leur coiffeur
Chapô	∅
Description	C'était une demande des salons de beauté. A partir du 1er janvier, les Californiens seront officiellement autorisés à siroter un verre d'alcool pendant qu'ils se font couper les cheveux ou tailler la barbe. A compter de 2017, les salons de beauté auront le droit, grâce à une nouvelle loi, de proposer jusqu'à 35 centilitres de bière -entre un demi et une pinte- ou un grand verre de vin à leurs clients sans avoir besoin de licence. Les clients se voyaient déjà depuis quelques années offrir du vin ou une bière dans cet Etat américain mais la pratique était souvent discrète, car interdite sans licence. Cette mesure, promulguée en septembre, n'est pas applaudie par tous, les associations de lutte contre l'alcoolisme craignant notamment qu'elle ne facilite l'accès des mineurs à l'alcool.

	Les salons de coiffure « ont toujours eu le droit de demander une licence s'ils voulaient servir de l'alcool », souligne Jim Kooler, d'un groupe de lutte contre la consommation d'alcool chez les jeunes, « the California Friday Night Live Partnership ».
Tags	Etats-Unis, alcool, californie, coiffure, salon
Date	2015-12-30
URL	http://www.20minutes.fr/monde/1987555-20161230-etats-unis-californiens-vont-bientot-pouvoir-trinquer-chez-coiffeur
Éditeur	20 Minutes

Article 6

Titre	Paris : violente rixe entre bandes rivales, un homme poignardé à mort
Chapô	∅
Description	Les choses ont dégénéré. Un jeune homme âgé de 29 ans est décédé mardi 21 au soir après avoir été grièvement blessé à coups de couteau lors d'une bagarre. Les faits se sont déroulés aux alentours de 19h30 à Paris. Pour une raison encore inconnue, une bagarre a éclaté entre des bandes rivales de deux quartiers : la Grange-aux-Belles pour le Xe arrondissement de la capitale et Chaufournier pour le XIXe arrondissement. Au total, une vingtaine d'individus ont pris part à la rixe et trois d'entre eux ont été blessés. Ils ont tous été conduits à l'hôpital Lariboisière. L'un d'eux a perdu la vie après avoir reçu quatre coups de couteau dont deux à l'abdomen. Alertés, les pompiers n'ont pas réussi à le réanimer malgré un massage cardiaque. Selon une source proche du dossier, qui s'est confiée à LCI, la victime était née en 1987, habitait le XIXe arrondissement et était défavorablement connu des services de police. Le deuxième homme, lui, a également été blessé à l'arme blanche. Quant au troisième, il a été frappé à la tête avec une barre de fer. Il souffre de blessures plus légères.

	Selon les dernières informations divulguées par Le Parisien, des renforts policiers ont été envoyés mardi soir pour sécuriser les abords de l'hôpital où des jeunes s'étaient regroupés "bouleversés d'apprendre la mort de leur camarade". L'un d'eux a d'ailleurs été interpellé pour rébellion. Pour éclaircir les circonstances de ce drame, une enquête a été ouverte. Elle a été confiée au 2e district de la police judiciaire.
Tags	Interpellation, Rébellion, Hôpital, Police, Bagarre, Paris, Policiers, Rixe, Couteau, Abdomen, Armes Blanches, Quartiers, Barre de fer, Mort, Blessés, Massage cardiaque, Blessures, Bandes, Individus, Capitale, Enquête
Date	2017-03-22
URL	http://www.francesoir.fr/societe-faits-divers/paris-violente-rixe-entre-bandes-rivales-un-homme-poignarde-mort-abdomen-bagarre-blesses-arrondissement
Éditeur	France Soir

Article 7

Titre	Ai-je raison d'être inquiète pour mon couple ?
Chapô	Mon mari et moi sommes tous les deux au chômage depuis plusieurs années. Nous nous soutenons et avons toujours été très complices. Mais, aujourd'hui, une amie ouvre une boutique et lui a proposé d'être son associé. Depuis, ils ne se quittent plus et, alors que j'attends notre deuxième enfant, je crains pour notre couple. Est-ce que je suis parano ? J'ai besoin d'un conseil objectif. Maellis, Nîmes
Description	Je ne crois pas, Maellis, que vous soyez « parano ». Voir son mari passer ses journées avec une autre n'est facile pour personne, et il est normal que vous vous posiez des questions. Je pense simplement qu'il y a deux pièges dans lesquels il ne faut pas tomber. Le premier est celui de la rivalité (femme à femme) avec cette femme. Et c'est un terrain sur lequel elle veut vous entraîner, puisqu'elle prend un malin plaisir à venir vous expliquer comment « fonctionne » votre mari. Façon de vous dire qu'elle le connaîtrait d'ores et déjà mieux que vous (tout le monde a le droit de rêver...).

Le second piège est celui dans lequel vous tomberiez si vous vous contentiez d'observer l'attitude de votre mari sans essayer de comprendre ce qu'il joue. Or, c'est essentiel. Car, si cette femme est importante pour lui, ce n'est pas tant à cause de ce qu'elle est qu'à cause de ce qu'elle représente pour lui. Depuis des années, en effet, il se sent – et c'est normal – diminué par le chômage et doute de sa valeur. Et elle surgit – en sauveur – pour lui proposer non seulement un emploi, mais une place de patron associé. Elle est donc, à ce titre, celle qui lui rend, par rapport à la vie sociale, une image restaurée de lui-même. Cela lui donne sur lui un pouvoir dont elle ne se prive sans doute pas de jouer, et il se méfie d'autant moins qu'elle est la femme de son meilleur ami.
Conclusion ? Ne pas jouer le rôle qu'elle attend que vous jouiez : celui de l'épouse revendicative et éplorée. Être patiente et attendre… en écoutant. Et en vous entourant d'ami(e)s qui vous aident à tenir. C'est important.

Tags ∅
Date 2016-04-12
URL http://www.psychologies.com/Couple/Crises-Divorce/Jalousie/Reponses-d-expert/Ai-je-raison-d-etre-inquiete-pour-mon-couple
Éditeur Psychologies

Article 8

Titre Colmar : Armés et encagoulés, ils braquent le livreur pour lui prendre ses deux pizzas
Chapô ∅
Description Livreur de pizzas, un métier à risque ? De toute évidence, il n'y a pas que les dangers de la route qui guettent les livreurs au coin de la rue. Vendredi soir, vers 22h, l'un d'entre eux, de la société Domino's à Colmar a été agressé par deux hommes cagoulés et armés qui ont exigé les clefs de son scooter… mais pas le fond de caisse, rapportent les Dernières Nouvelles d'Alsace.

	Le livreur a fini tout de même par obtenir que ses agresseurs lui redonnent les clés du scooter et n'ont pris la fuite qu'avec les pizzas, pour un montant de 32,40 euros précisent nos confrères. Il a déposé plainte pour vol aggravé. Selon les premières informations, la personne qui avait commandé les pizzas ne les a jamais réclamées, ce qui laisse à penser a un guet-apens.
Tags	pizza, agression, Armes, scooter, colmar
Date	2016-12-12
URL	http://www.20minutes.fr/strasbourg/1978691-20161212-colmar-armes-encagoules-braquent-livreur-prendre-deux-pizzas
Éditeur	20 Minutes

Article 9

Titre	Recettes inratables
Chapô	∅
Description	Je vous livre là les secrets de cuisine de ma grand-mère, faites-en bon usage et surtout partagez-les tout autour de vous ! Comment réussir béchamel, beignets, cake, caramel, civet de lapin... Cette astuce vaut aussi pour les autres sauces. En fait il existe 2 astuces pour réussir à coup sûr béchamel ou tout autre sauce à base de farine. Mais tout d'abord commençons par la préparation du roux. Et laissez-moi vous livrer le 1er secret de la béchamel. (Le contraste de température) – Faites fondre le beurre dans une casserole puis ajoutez progressivement la farine. L'ensemble doit cuire à feu très doux pendant 5 minutes. Le roux est prêt et chaud. – Incorporez le lait froid (impérativement froid). Si vous avez réalisé le roux la veille et qu'il sort du réfrigérateur, c'est au contraire du lait bouillant qu'il faudra verser. Mais il existe un 2ème secret ! – Mélanger sa sauce (agrémentée d'un bouquet garni) avec une fourchette piquant une demie pomme de terre crue. Vous être en train de faire un cake et vous voulez que les raisins secs se répartissent uniformément dans la pâte ? Alors c'est tout simple, roulez-les dans du sucre glace ou de la farine avant de les incorporer.

	Impossible de rater votre caramel si vous suivez bien ces proportions : – 5 morceaux de sucre pour 1 ciil. à café d'eau. – Dès qu'il commence à brunir, ajouter 1 filet de jus de citron, voire quelques gouttes de vinaigre, faute de quoi vous pourrez dire adieu à votre casserole Une cuillerée à café de levure chimique dans la pâte rendra les beignets beaucoup plus légers. Une astuce de grand-chef ! Pour parfumer la sauce de votre civet de lapin, rajoutez 3 bonnes cuillères à soupe de café très fort dans le vin de cuisson.
Tags	Cuisine, Astuces de cuisine
Date	2017-05-16
URL	https://www.remedes-de-grand-mere.com/remede/recettes-inratables/
Éditeur	Remèdes de Grand-Mère

Article 10

Titre	Washington menace Pyongyang d'une "réponse militaire massive"
Chapô	∅
Description	Le président américain Donald Trump a dénoncé dimanche l'essai nucléaire "hostile" mené par la Corée du Nord, et le Pentagone a promis à Pyongyang une "réponse militaire massive" en cas de "menace" visant les Etats-Unis. Peu après le test de cet engin qui était selon Pyongyang une bombe à hydrogène ou bombe H - "une réussite parfaite", selon la télévision publique nord-coréenne -, le général Jim Mattis, secrétaire américain à la Défense, est monté en première ligne. Depuis la Maison Blanche, il a adressé une mise en garde solennelle au régime de Kim Jong-Un, tout en l'appelant à entendre les injonctions de la communauté internationale. "Nous avons de nombreuses options militaires et le président voulait être informé sur chacune d'entre elles", a-t-il lancé à l'issue d'une réunion avec Donald Trump et son équipe de sécurité nationale. "Toute menace visant les Etats-Unis ou ses territoires, y compris Guam (dans le Pacifique, ndlr), ou ses alliés, fera l'objet d'une réponse militaire massive", a-t-il averti, le général Joe Dunford, chef d'état-major interarmées des Etats-Unis, debout à ses côtés.

M. Mattis a aussi précisé que les Etat-Unis ne cherchaient en aucune manière "l'anéantissement total" de la Corée du Nord et a appelé Pyongyang à prêter attention aux mises en garde du Conseil de sécurité de l'ONU.
Le Conseil, qui a déjà infligé en vain sept trains de sanctions à Pyongyang pour le contraindre à renoncer à ses ambitions nucléaires et balistiques, doit se réunir lundi matin en urgence.
Quelques heures auparavant, le régime nord-coréen avait publié des photos montrant son dirigeant Kim Jong-Un en train d'inspecter un engin présenté comme une bombe H (bombe à hydrogène ou thermonucléaire) pouvant être installée sur le nouveau missile balistique intercontinental dont dispose le régime.
- 'Nous verrons' -
Donald Trump, qui a dénoncé "des actions dangereuses pour les Etats-Unis", a laissé planer le doute sur ses intentions. A un journaliste qui lui demandait à la sortie d'une église, en cette journée nationale de prière pour les victimes de la tempête Harvey, s'il envisageait une réponse militaire, le président américain a répondu : "Nous verrons".
"La Corée du Sud s'aperçoit, comme je le leur ai dit, que leur discours d'apaisement avec la Corée du Nord ne fonctionnera pas, ils ne comprennent qu'une chose !", a lancé M. Trump sur Twitter à l'intention de son homologue sud-coréen Moon Jae-In, partisan d'un dialogue avec le régime de Kim Jong-Un.
Les experts estiment que l'option militaire contre le régime de Kim Jong-Un est extrêmement risquée, car elle pourrait provoquer une réaction en chaîne et un grave conflit régional.
La Corée du Sud a mené dimanche soir un exercice militaire impliquant des missiles balistiques en réponse à l'essai nucléaire nord-coréen, a rapporté l'agence de presse sud-coréenne Yonhap.
L'exercice a simulé une attaque sur le polygone d'essais nord-coréen, touchant "des cibles choisies dans la mer de l'Est" ou mer du Japon, a indiqué Yonhap, citant l'état-major interarmes.
Alors que le secrétaire américain au Trésor Steven Mnuchin évoquait de possibles nouvelles sanctions, le président Trump a brandi le menace d'arrêter "tous les échanges commerciaux" avec "tout pays faisant des affaires avec la Corée du Nord".
Le président n'a donné aucune précision sur cette menace qui apparaît impossible à appliquer à la lettre : la Chine, partenaire économique central des Etats-Unis, est destinataire de quelque 90% des exportations nord-coréennes.

Les Etats-Unis ont également commencé à frapper de sanctions des entités chinoises et russes qui ont des relations d'affaires avec la Corée du Nord. Mais l'exercice est politiquement délicat vis-à-vis de Pékin.
Le secrétaire général de l'ONU, Antonio Guterres, a condamné l'essai nucléaire, qu'il a qualifié de "profondément déstabilisant".
Le président français Emmanuel Macron et la chancelière allemande Angela Merkel ont prôné des sanctions bilatérales de l'Union européenne en plus de celles qui pourraient être décidées à l'ONU.
- Ordre manuscrit de Kim -
Principal allié de Pyongyang, la Chine a "condamné vigoureusement" la nouvelle provocation nord-coréenne. Elle a également entrepris des contrôles de radiations nucléaires à sa frontière avec la Corée du Nord.
Pyongyang n'a jamais caché que ses programmes interdits avaient pour but de mettre au point des missiles balistiques intercontinentaux susceptibles de porter le feu nucléaire sur le continent américain.
La télévision d'Etat nord-coréenne a diffusé une image de l'ordre manuscrit de Kim Jong-Un demandant que l'essai soit conduit ce 3 septembre à midi heure locale. M. Kim a souligné que "tous les composants de cette bombe H ont été fabriqués à 100% nationalement", selon l'agence de presse officielle nord-coréenne KCNA.
Selon des spécialistes sud-coréens, la puissance de la nouvelle secousse était cinq à six fois supérieure à celle du précédent essai nord-coréen, effectué en septembre 2016 et qui était de 10 kilotonnes.
Quelle que soit la puissance de la déflagration, Jeffrey Lewis, du site armscontrolwonk.com, a estimé qu'il s'agissait bien d'une arme thermonucléaire, ce qui constitue un progrès notoire dans les programmes nucléaire et balistique nord-coréens.
Pour Koo Kab-Woo, spécialiste de la Corée du Nord à l'Université de Seoul, "la Corée du Nord continuera avec son programme d'armes nucléaires à moins que les Etats-Unis ne proposent des discussions".

Tags ∅
Date 2017-09-04
URL http://www.notretemps.com/accueil/washington-menace-pyongyang-d-une-afp-201709,i149777
Éditeur Notre Temps

Annexe B

Vidéos : Exemples du corpus

Cette annexe présente 10 exemples de vidéos de la colletion de Mediabong, reflétant la variété de leurs contenus :

Vidéo 1

Titre	Un documentaire pour raconter l'histoire du Vietnam
Description	A partir de mardi et pour trois soirées de suite, "Arte" propose neuf heures de documentaire en prime time pour raconter l'histoire du Vietnam. Il a fallu plus de dix ans de travail aux deux réalisateurs pour réaliser ce documentaire.
Tags	eva roque, medias, video
Date	2017-09-19
Producteur	Europe 1

Vidéo 2

Titre	Attentats en Catalogne : la cellule de Ripoll a priori démantelée
Description	Au sommaire de cette deuxième partie : les attentats en Espagne et la traque des jihadistes. La cellule de Ripoll a, à priori, été démantelée. Aux États-Unis, Donald Trump est contraint à un nouveau revirement : les troupes américaines resteront finalement en Afghanistan.
Tags	∅
Date	2017-08-25
Producteur	France 24

Vidéo 3

Titre	Avion russe : l'escalade se poursuit entre Moscou et Ankara

Description L'escalade verbale se poursuit entre Moscou et Ankara. Ce mardi, la Turquie a abattu un avion russe à la frontière syrienne, l'accusant d'avoir violé son espace aérien. L'armée turque a diffusé un enregistrement, présenté comme les sommations qui ont précédé le tir. Mais le Kremlin dément, en livrant le témoignage d'un homme présenté comme l'un des pilotes du chasseur, et durcit le ton. 'Nous n'avons entendu aucune excuse de la part des dignitaires politiques turcs', affirme le Président russe Vladimir Poutine, 'ni de proposition de compensation des dommages, ni de promesse de punir les criminels. Cela donne l'impression que les dirigeants turcs mènent volontairement les relations russo-turques droit dans l'impasse. Nous le regrettons'. 'Si la même violation se produisait aujourd'hui, la Turquie réagirait de la même manière', indique le chef de l'Etat turc Recep Tayyip Erdogan. 'Ce n'est pas le pays dont l'espace est violé, mais celui qui procède à la violation qui devrait s'interroger et prendre les mesures nécessaires pour éviter qu'un tel incident ne se reproduise'. La Russie a renforcé son arsenal militaire à Lattaquié, en Syrie, et la Turquie a fait de même près de la frontière syrienne. Les deux puissances s'accusent mutuellement de contribuer à l'essor du groupe Etat Islamique. 'Les terroristes et leur trafic illégal de pétrole, d'êtres humains, n'étaient pas seulement couverts, ils le sont toujours. Mais certains se font même de l'argent là-dessus. Des centaines de millions, voire des milliards de dollars', souligne Vladimir Poutine. 'Si vous cherchez les responsables de l'armement et du soutien financier de Daech,vous devriez regarder en premier du côté du régime de Bachar Al-Assad', rétorque Recep Tayyip Erdogan, 'et des pays qui le soutiennent, qui agissent avec lui'. Le Premier ministre russe Dmitri Medvedev a demandé à son gouvernement des mesures pour geler ses investissements en Turquie. La Russie a appelé ses ressortissants à quitter la Turquie. Le Président turc, Recep Tayyip Erdogan, a annoncé de son côté que les projets avec la Russie pourraient être annulés. Ce mercredi, le Premier ministre islamo-conservateur Ahmet Davutoglu avait pourtant déclaré que son pays n'avait pas l'intention d'aggraver les tensions avec la Russie, 'pays ami et voisin', qui fournit à Ankara près de la moitié de sa consommation de gaz naturel.

Tags Catastrophe aérienne, Recep Tayyip Erdogan, Russie, Tension diplomatique, Turquie, Vladimir Poutine

Date 2015-11-25

Producteur Euronews

Vidéo 4

Titre	Kristen Stewart et Stella Maxwell : Elles s'affichent enfin publiquement !
Description	Même si Kristen Stewart et Stella Maxwell ne sont pas prêtes à s'embrasser en public, elles acceptent maintenant de se faire photographier ensemble sans se cacher.
Tags	STAR24, TV, 100%, People, kristen stewart, stella maxwell, couple, celibat, love, amour, officiel, cinema, mannequin, soko, robert pattinson, victoria secret
Date	2017-01-06
Producteur	Star24TV

Vidéo 5

Titre	Amazon va créer 1 500 emplois mais les ESN françaises, plus discrètes, créent aussi des milliers de postes - 25/02
Description	Frédéric Simottel revient sur l'actualité tech de la semaine. Au sommaire de cette édition : Amazon va créer 1 500 emplois mais les ESN françaises, plus discrètes, créent aussi des milliers de postes. L'intelligence artificielle ne remplacera pas les hommes, selon Paul Hermelin, PDG de Capgemini. Le marché français des data centers en croissance : +20% en surface d'ici 2020. Les robots vont-ils bientôt cuisiner à notre place ? Projet Aadhaar : 1,2 milliards d'Indiens sont désormais identifiables par l'iris et les empreintes digitales. - 01 Business Forum - L'hebdo, du samedi 25 février 2017, présenté par Frédéric Simottel, sur BFM Business.
Tags	projet aadhaar, identite numer, création, Amazon, Frédéric Simottel, robots, intelligence artificielle, fiches, Croissance, Poste, Iris, franceeco, empreintes digitales, emission, 20%, cuisine, data centers, 2020, replay, la revue de presse, scan, emploi, tech, 01 business, ESN
Date	2017-02-25
Producteur	BFM Business

Vidéo 6

Titre	Lors de l'élection de Miss America, Miss Texas attaque violemment Donald Trump !
Description	∅

Tags	∅
Date	2017-09-17
Producteur	Tribunal du Net

Vidéo 7

Titre	Mercedes classe b 220 occasion visible à Saint loup de varennes présentée par Mercedes chalon
Description	Cette Mercedes classe b 220 occasion disponible à Saint loup de varennes dans le 71 est mise en vente par Mercedes chalon - 03 45 58 09 26 - tous les détails sont sur http://www.auto-selection.com/voiture-occasion/mercedes-classe-b-220/14232422.html?utm_source=youtube&utm_medium=youtube&utm_campaign=youtube
Tags	Cette, Mercedes, classe, 220, occasion, disponible, Saint, loup, varennes, dans, est, mise, vente, par, Mercedes, chalon, tous, les, détails, sont, sur, http://www.auto-selection.com/voiture-occasion/mercedes-classe-b-220/14232422.html?utm_source=youtube&utm_medium=youtube&utm_campaign=youtube, b, à, de, le, 71, en, -, 03, 45, 58, 09, 26, -
Date	2016-09-12
Producteur	Auto Selection

Vidéo 8

Titre	À l'ONU, Macron s'oppose à Trump
Description	Le président de la République a pris la parole ce mardi soir à l'ONU. L'occasion pour lui de défendre sa vision du monde, en contradiction avec celle de Donald Trump.
Tags	∅
Date	2017-09-19
Producteur	L'Express

Vidéo 9

Titre	Recette de quiche thon et tomate - Ptitchef.com
Description	Pour une recette simple et efficace, rien de mieux qu'une bonne QUICHE THON ET TOMATE ! Déroulez ci-dessous pour en savoir davantage Lorsqu'on cherche une recette maison rapide à faire et bonne à manger, la quiche s'impose comme une évidence ;-) Sortez vos pâte brisée et c'est parti pour la réalisation ! La vidéo vous a plu mais vous souhaitez plus de détails sur la recette ? Découvrez-la en version écrite illustrée de photos dans le lien suivant : https://www.ptitchef.com/recettes/plat/quiche-au-thon-et-a-la-tomate-fid-1568866 Et n'oubliez que de nouvelles recettes vous attendant chaque jour sur notre site : https://www.ptitchef.com/ Informations sur la musique de la vidéo : Source : https://www.youtube
Tags	ptitchef, petitchef, recette, cuisine, vidéo, quiche, thon, tomate, facile, recette facile, recette rapide, rapide, express, quiche maison, maison, fait maison, quiche facile, tuto, tutoriel, tutorial, pas à pas, stop motion, recipe, cook, cooking, food, tomato, fish, tuna, easy recipe, easy, home made, home made recipe, diy, diy recipe, diy cooking, express recipe, fast, fast recipe, pique nique
Date	2017-08-28
Producteur	Ptitchef

Vidéo 10

Titre	Ligue 1 - Premiers nuages dans le ciel parisien
Description	Avec l'imbroglio entre Cavani et Neymar ainsi que l'affaire Kurzawa, le Paris Saint-Germain se confronte à de réels problèmes pour la première fois de la saison alors que le club brille sur le plan sportif.
Tags	Neymar, football, RMCSagence, Ligue 1, conflits, cavani, PSG, Affaire kurzawa, tensions, Paris Saint-Germain
Date	2017-09-19
Producteur	RMC Sport

Annexe C

Mediabong : nomenclature thématique

Cette annexe présente la liste des 24 thèmes et 131 sous-thèmes de la nomenclature Mediabong :

1. MAISON
   - Deco et design
   - Jardinage
   - Autres - Maison
   - Astuces pratiques
   - Bricolage
2. SPORT
   - American football
   - Baseball
   - Tennis
   - Turf
   - Sports extrêmes
   - Basket
   - Sport équestre
   - Sports d'hiver
   - Cyclisme
   - Autres sports
   - Sports mécaniques
   - Rugby
   - Sports de combat
   - Football
   - Voile
3. INTERNATIONAL
   - Int>Canada
   - Autres - International
   - Int>France BE

- Int>Ameriques
- Int>Afrique et moyen-orient
- Int>Europe
- Int>Asie-Océanie

4. POLITIQUE
5. SANTE
   - Sexo
   - Fitness
   - Régime
   - Psycho
   - Bien-etre
   - Maladies
   - News santé
   - Autres - Santé
   - Grossesse et enfant
6. HUMOUR
7. TOURISME VOYAGE
   - Allemagne
   - Océanie
   - Moyen-Orient
   - Afrique
   - USA - Canada
   - Asie
   - Amérique du Sud - Centrale
   - France
   - Europe
   - Autres Tourisme-voyage
8. BEAUTE
   - Beauté - Homme
   - Coiffure
   - Parfum
   - Autres - Beauté
   - Maquillage
   - Soins
9. MODE
   - Autres - Mode
   - Mode Femme
   - Mode Homme

- Accessoires

10. HIGH TECH
    - Web et appli
    - Mobile et tablette
    - Logiciel
    - Hardware
    - Apple
    - Gaming et console
    - Photo, video, son
    - Autres - High-tech
11. PEOPLE
    - People Belgique
    - People Allemagne
    - People France
    - People International
12. ECONOMIE - FINANCE
    - Droit
    - Assurance
    - Immobilier
    - Entreprise
    - Emploi
    - Bourse et marches
    - Impôt, retraite, patrimoine
    - Autres - Eco-Finances
13. AUTO / MOTO
    - Autres - Auto/moto
    - News moto
    - Essai moto
    - News auto
    - Essai auto
14. ART ET CULTURE
    - Histoire
    - Architecture
    - Theatre-Spectacles
    - Autres - Art-culture
    - Musée
    - Peinture et Sculpture
    - Littérature

15. ECOLOGIE - ENVIRONNEMENT
    - Déchets et recyclage
    - Pollution
    - Autres- Ecologie
    - Protection de l environnement
    - Climat
    - Energie
16. CUISINE
    - Vins - boissons
    - Recettes-Fêtes
    - Restaurants
    - Recettes- Pâtes
    - Recettes-Entrées
    - Recettes-Desserts
    - Autres - Cuisine
    - Recettes- Sauces
    - Recettes-Poissons
    - Recettes-Legumes
    - Recettes du Monde
    - Recettes-viandes
    - Recettes-soupes
17. ANIMAUX
    - Chevaux
    - Autres - Animaux
    - Chats
    - Chiens
18. SCIENCES
19. INSOLITE
20. ACTU MEDIA
21. ACTUALITE
22. MUSIQUE
    - Autres - Musique
    - Variete Belge
    - Rap, RnB
    - Variété française
    - Variété internationale
23. CINEMA
    - Science fiction

- Drame
- Comédie
- Animation
- Action
- Autres - Cinéma
- Séries

24. SOCIETE
    - Education - Formation
    - Emploi, consommation
    - Loisirs
    - Jeunesse
    - Justice - faits divers
    - Autres - Société
    - Transports

Annexe D

Ensemble de test : liste des requêtes

Cette annexe présente la liste des 50 articles constituant les requêtes de l'ensemble de test. Dans un souci de lisibilité, nous n'en présentons ici que les titres, mais c'est bien à partir de l'intégralité du contenu textuel des articles que les calculs se font.

1. Pas de pause des enquêtes sur des candidats durant la campagne pour Urvoas - Autres - Notre Temps
2. Bercy : Quand Macron dépensait 120 000 euros en 8 mois pour ses repas en bonne compagnie
3. Si vous utilisez Facebook, vous vivrez potentiellement plus longtemps
4. Otan : Washington menace de "modérer son engagement" - Autres - Notre Temps
5. Le bureau des jardins et des étangs : une histoire originale, belle et très bien écrite
6. La recette du chia pudding d'Olivia
7. Sexe, tourments et libido : une hotline pour les jeunes Afghans - Autres - Notre Temps
8. César : Huppert, Cotillard, Ulliel, Sy parmi les nommés - Autres - Notre Temps
9. Discrétion et sobriété pour les funérailles de Fidel Castro
10. "Quartiers populaires" ? Mais que de mensonges derrière ces mots !
11. VIDEO. Usain Bolt s'est entraîné avec une équipe de DH de la Côte d'Azur (mais il est pas terrible)
12. Législatives : mécontent du candidat investi par Les Républicains, Patrick Balkany fait retirer les affiches de François Fillon à Levallois
13. L'administration Trump en difficulté sur ses relations avec Moscou - Autres - Notre Temps
14. Entreprises : Pourquoi les élections professionnelles dans les TPE sont importantes
15. Grippe aviaire : la justice ouvre une enquête après le scoop sur l'abattage des canards du Sud-Ouest
16. A Nancy, l'Eglise protestante unie assume son ouverture aux mariés homosexuels
17. Syrie : L'ONU n'identifie pas les responsables de l'attaque d'un convoi humanitaire

18. Londres : publication du projet de loi sur le déclenchement du Brexit - Autres - Notre Temps
19. Nabilla va-t-elle jouer dans la série «Orange is the New Black» ?
20. Bolloré visé par une enquête en Italie dans l'affaire Mediaset - Autres - Notre Temps
21. Haute couture à Paris : Schiaparelli pop, Iris Van Herpen organique - Autres - Notre Temps
22. Patricia Kaas revient avec un nouvel album et lance une tournée européenne
23. Télétravail : le secteur public est-il au-dessus du droit du travail ?
24. Miss France : Les rêves d'Alicia ? Une famille «soudée», du mannequinat et «Touche pas à mon poste»
25. Galileo : plusieurs horloges atomiques en panne - Autres - Notre Temps
26. Strasbourg : Proprios qui louent dans l'illégalité des appartements « type Airbnb », c'est la saison des contrôles
27. Violence dans les abattoirs : les députés refusent les caméras
28. Le Nigeria ouvre une enquête après un bombardement qui a fait 70 morts - Autres - Notre Temps
29. VIDEO. Clasico : Le coup de boule de Sergio Ramos qui offre sur le fil le nul au Real Madrid
30. Corse : Ils veulent rencontrer Christophe Maé, leur message cartonne sur les réseaux sociaux
31. Lille : Gérard Lopez nouveau propriétaire - Autres - Notre Temps
32. Intempéries : fin de l'alerte orange, 23.000 foyers sans électricité - Autres - Notre Temps
33. Moteurs diesel : des juges d'instruction vont enquêter sur Renault - Autres - Notre Temps
34. Quand soins et parfums nous coachent : La beauté, outil de développement personnel
35. Derrière l'alliance avec Emmanuel Macron, le projet non-exprimé de François Bayrou de recréer l'UDF
36. 5 idées reçues sur la dépression : Idée reçue #4 : La dépression est un problème de riches
37. Russie : des héros de la 2e Guerre qui n'ont sans doute jamais existé - Autres - Notre Temps

38. Un militaire de l'opération Sentinelle patrouillant au Louvre attaqué au couteau se défend et neutralise son agresseur
39. Marks and Spencer : le CE va agir en justice pour défendre les emplois - Retraite - Notre Temps
40. A Alep, les civils doivent aussi faire face à la soif, à la faim et au froid
41. Présidentielle 2017 : Fillon veut «un audit des comptes sociaux» par «des experts indépendants»
42. Attentat manqué du Thalys : Le djihadiste affirme qu'il ciblait des Américains
43. En février : bouturer le syngonium
44. VIDEO : Des événements de 2016, ils font un film d'horreur parodique
45. Montpellier : une journaliste frappée par des lycéens lors d'une manifestation pro-Théo
46. Au Ritz, "maison" de Chanel, Lagerfeld célèbre les métiers d'art - Autres - Notre Temps
47. Crash du vol EgyptAir : L'Egypte commence à rendre aux familles les dépouilles des victimes
48. Famine au Soudan du Sud : le président promet d'aider l'accès aux ONG - Autres - Notre Temps
49. Les filles de Jacqueline Sauvage ont déposé une demande de grâce totale à l'Elysée
50. Brexit : May souhaite une sortie aussi "en douceur que possible" - Autres - Notre Temps

Annexe E

Extraction d'expressions multi-mots : exemple de sortie

Cette annexe présente un exemple de sortie fournie par l'algorithme d'extraction d'EMM, à partir d'un corpus d'articles récupérés entre le le 26 juin et le 2 juillet 2017.

La première colonne présente l'ensemble des candidats extraits par l'algorithme, et les deux suivantes correspondent aux annotations manuelles associées à ces candidats. La deuxième colonne correspond à l'accord ou non pour l'intégration du terme à la ressource terminologique. La troisième colonne correspond au type associé au terme, lorsque celui-ci a été retenu.

TERME CANDIDAT	INTÉGRATION RESSOURCE	TYPE ASSOCIÉ
république en marche	OUI	Divers
sylvie goulard	OUI	Personne
richard ferrand	OUI	Personne
fourgon de gendarmerie	NON	–
huile de coco	OUI	Divers
jacqueline jacob	OUI	Personne
murielle bolle	OUI	Personne
françois de rugy	OUI	Personne
abus de position dominante	OUI	Divers
appel d'air	OUI	Divers
véronique robert	OUI	Personne
époux jacob	NON	–
présidence du groupe	NON	–

Annexe F

Semiabong : exemples d'appariements automatiques

Cette annexe présente un échantillon de 45 appariements automatiques proposés par Semiabong, sur des articles traités entre le 3 et le 9 juillet 2017. Les titres des articles et des vidéos appariées y sont présentés, ainsi que leurs dates respectives de publication.

Article	-	François de Rugy élu président de l'Assemblée nationale	-	2017-07-03
Vidéo	-	François de Rugy élu président de l'Assemblée nationale	-	2017-06-27
Article	-	Le Comité d'éthique dit oui à un élargissement du cadre de la PMA	-	2017-07-03
Vidéo	-	PMA "pour toutes les femmes" : qu'implique l'avis du Comité d'éthique ?	-	2017-06-28
Article	-	Migrants : Béatrice Huret reconnue coupable mais dispensée de peine	-	2017-07-03
Vidéo	-	Cette ancienne militante FN est jugée pour avoir aidé un migrant à traverser la Manche	-	2017-06-27
Article	-	Macron devant le Congrès pour fixer le cap de son quinquennat	-	2017-07-03
Vidéo	-	Les précédents Congrès de Versailles avant celui d'Emmanuel Macron lundi	-	2017-07-02
Article	-	Washington lève l'interdiction des ordinateurs pour les vols d'Etihad	-	2017-07-03
Vidéo	-	Sécurité : les ordinateurs interdits dans les vols à destination des Etats-Unis ?	-	2017-05-28

Article	-	L'argentier du Vatican inculpé pour pédophilie clame son innocence	-	2017-07-04
Vidéo	-	L'argentier du Vatican inculpé pour pédophilie	-	2017-06-29
Article	-	Allemagne : le Parlement autorise le mariage homosexuel	-	2017-07-04
Vidéo	-	Le Bundestag fête le mariage pour tous	-	2017-06-30
Article	-	Pyongyang tire un nouveau missile, Trump dénonce une "absurdité"	-	2017-07-04
Vidéo	-	La Corée du Nord procède à un nouveau tir de missile (et le rate)	-	2017-04-29
Article	-	A Mossoul, l'EI multiplie les attentats suicide pour freiner les forces irakiennes	-	2017-07-04
Vidéo	-	À Mossoul, au plus près des combats, avec les snipers des Forces spéciales irakiennes	-	2017-06-29
Article	-	Se baigner dans la Seine, un rêve bientôt réalité ?	-	2017-07-04
Vidéo	-	On peut nager dans le bassin de la Villette, en attendant de se baigner (un jour ?) dans la Seine	-	2017-06-17
Article	-	G20 sous haute tension en vue avec Donald Trump	-	2017-07-05
Vidéo	-	Retour de veste, Donald Trump reconduit la levée des sanctions iraniennes	-	2017-07-04
Article	-	Une start-up auvergnate réconcilie disque vinyle et numérique	-	2017-07-05
Vidéo	-	"Le vinyle est à la musique ce que la chair est au désir"	-	2017-07-03
Article	-	PHOTOS Obsèques nationales de Simone Veil : des personnalités lui rendent hommage	-	2017-07-05
Vidéo	-	La cérémonie obsèques Simone Veil	-	2017-07-05
Article	-	Un journaliste du "Canard enchaîné" : "Sur Fillon, on en a encore sous le pied"	-	2017-07-05
Vidéo	-	"La vérité, c'est que ça fait rire tout le monde" : le rédacteur en chef du Canard enchaîné s'exprime sur l'affaire Fillon	-	2017-01-31

Article	-	La Chine invite des cancérologues étrangers à soigner le dissident Liu Xiaobo	-	2017-07-05
Vidéo	-	Sous pression, la Chine invite des médecins étrangers au chevet de Liu Xiaobo	-	2017-07-05
Article	-	Venezuela : des pro-Maduro séquestrent les députés, 7 parlementaires blessés	-	2017-07-06
Vidéo	-	Venezuela : assaut de civils armés au parlement	-	2017-07-05
Article	-	Okja : l'avis enthousiaste d'une jeune fille de 12 ans	-	2017-07-06
Vidéo	-	Cannes 2017 : «Okja», de Bong Joon-ho	-	2017-05-19
Article	-	France : les baisses d'impôts seront décidées "avant la fin de l'année"	-	2017-07-06
Vidéo	-	Réformes fiscales : la baisse des impôts pas pour tout de suite... mais la hausse de la CSG si !	-	2017-07-05
Article	-	Fac : sélection plutôt que tirage au sort ?	-	2017-07-06
Vidéo	-	Admission Post-Bac : encore 117.000 lycéens sans affectation après la deuxième phase	-	2017-06-30
Article	-	Vaccins obligatoires : le gouvernement travaille à une clause d'exemption	-	2017-07-06
Vidéo	-	Santé : de trois à onze vaccins obligatoires pour les enfants ?	-	2017-06-16
Article	-	La guerre continue entre Apple et Qualcomm, qui demande l'arrêt des importations d'iPhones	-	2017-07-07
Vidéo	-	01LIVE HEBDO #138 Apple - Qualcomm : jusqu'où peut aller l'affrontement ?	-	2017-04-13
Article	-	Une "Antigone" entre Orient et Occident en ouverture du festival d'Avignon	-	2017-07-07
Vidéo	-	Top départ pour le 71e Festival d'Avignon	-	2017-07-07
Article	-	Syrie : au moins 18 morts dans l'attentat suicide de Damas	-	2017-07-07
Vidéo	-	Syrie : attentat suicide à Damas	-	2017-07-02

Article	-	Migrants : opération d'évacuation de campements dans le nord de Paris	-	2017-07-07
Vidéo	-	Evacuation des migrants à La Chapelle	-	2017-07-07
Article	-	L'ex-candidat à la présidentielle, Benoît Hamon quitte le Parti socialiste	-	2017-07-07
Vidéo	-	Benoît Hamon quitte le parti socialiste et crée son "mouvement du 1er juillet"	-	2017-07-01
Article	-	Canada : excuses et dédommagement pour un ex-détenu de Guantanamo	-	2017-07-08
Vidéo	-	Le Canada dédommage un ex-détenu de Guantanamo	-	2017-07-05
Article	-	Epilogue pour Trump contre le reste du monde au G20	-	2017-07-08
Vidéo	-	G20 : tous contre Trump ?	-	2017-07-07
Article	-	Cancers du sein : le docétaxel peut à nouveau être utilisé	-	2017-07-08
Vidéo	-	Cancer du sein : une enquête ouverte sur le docétaxel après la mort de 5 patientes	-	2017-02-16
Article	-	Premiers bouchons sur la route des vacances	-	2017-07-08
Vidéo	-	Bison Futé prévoit un weekend chargé sur les routes	-	2017-07-07
Article	-	Coupe des Confédérations : finale Chili-Allemagne, à qui la coupe ?	-	2017-07-08
Vidéo	-	Coupe des Confédérations – Di Meco : "Le Chili est favori pour cette finale"	-	2017-07-01
Article	-	En quête de renouvellement, le PS change de tête	-	2017-07-09
Vidéo	-	Conseil national du Parti socialiste : formation de la direction collégiale	-	2017-07-08
Article	-	Le bitcoin, une monnaie virtuelle qui intéresse malgré les risques	-	2017-07-09
Vidéo	-	Bitcoin, comment ça marche ? Est-ce fiable ?	-	2017-07-07
Article	-	Sexisme dans la Silicon Valley : les langues se délient, les scandales se multiplient	-	2017-07-09

Vidéo	-	What's Up New York : Comment la Silicon Valley gère-t-elle les cas de harcèlement sexuel ?	-	2017-07-07
Article	-	JO-2024/2028 : le CIO entame la semaine du qui perd gagne	-	2017-07-09
Vidéo	-	Paris 2024 – Le rapport de la commission d'évaluation du CIO	-	2017-07-06
Article	-	Après Paris, Berlin : marche d'imams contre le terrorisme	-	2017-07-09
Vidéo	-	Allemagne : des musulmans contre le terrorisme	-	2017-06-17

Annexe G

Semiabong : Formule de pondération de termes

Cette annexe est soumise à confidentialité et nécessite une demande de consultation par mail, à l'adresse adele.desoyer@gmail.com.

Bibliographie

Abberley, Dave, Kirby, David, Renals, Steve et Robinson, Tony (1999). « The THISL broadcast news retrieval system ». *in* : *ESCA Tutorial and Research Workshop (ETRW) on Accessing Information in Spoken Audio*.

Abeillé, Anne, Clément, Lionel et Toussenel, François (2003). « Building a treebank for French ». *in* : *Treebanks*, p. 165–187.

Adam, Jean-Michel (1997). « Unités rédactionnelles et genres discursifs : cadre général pour une approche de la presse écrite ». *in* : *Pratiques* **94**, p. 3–18.

Adomavicius, Gediminas, Manouselis, Nikos et Kwon, YoungOk (2015). « Multi-criteria recommender systems ». *in* : *Recommender systems handbook*. Springer, p. 847–880.

Adomavicius, Gediminas et Tuzhilin, Alexander (2005). « Toward the next generation of recommender systems : A survey of the state-of-the-art and possible extensions ». *in* : *IEEE transactions on knowledge and data engineering* **17**.*6*, p. 734–749.

Allan, James (2002). « Introduction to topic detection and tracking ». *in* : *Topic detection and tracking*. Springer, p. 1–16.

Allan, James, Carbonell, Jaime G., Doddington, George R., Yamron, Jonathan et Yang, Yiming (1998). « Topic detection and tracking pilot study final report ». *in* : *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*, p. 194–218.

Allan, James, Lavrenko, Victor et Nallapati, Ramesh (2002). « UMass at TDT 2002 ». *in* : *Topic Detection and Tracking : Workshop*.

Allan, James, Papka, Ron et Lavrenko, Victor (1998). « On-line new event detection and tracking ». *in* : *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, p. 37–45.

Amatriain, Xavier (2013). « Mining large streams of user data for personalized recommendations ». *in* : *ACM SIGKDD Explorations Newsletter* **14**.*2*, p. 37–48.

Amini, Massih-Reza et Gaussier, Eric (2013). *Recherche d'information : applications, modèles et algorithmes*. Editions Eyrolles.

Ariki, Yasuo et Sugiyama, Yoshiaki (1997). « A TV news retrieval system with interactive query function ». *in* : *Cooperative Information Systems, 1997. COOPIS'97., Proceedings of the Second IFCIS International Conference on*. IEEE, p. 184–192.

Aslam, Javed, Diaz, Fernando, Ekstrand-Abueg, Matthew, McCreadie, Richard, Pavlu, Virgil et Sakai, Tetsuya (2014). « TREC 2013 Temporal Summarization ». *in* : *Proceedings of the 22th Text REtrieval Conference (TREC2013)*.

Aslam, Javed, Diaz, Fernando, Ekstrand-Abueg, Matthew, McCreadie, Richard, Pavlu, Virgil et Sakai, Tetsuya (2015). « TREC 2014 temporal summarization track overview ». *in* : *Proceedings of the 23th Text REtrieval Conference (TREC2014)*.

Aslam, Javed, Diaz, Fernando, Ekstrand-Abueg, Matthew, McCreadie, Richard, Pavlu, Virgil et Sakai, Tetsuya (2016). « TREC 2015 temporal summarization track overview ». *in* : *Proceedings of the 24th Text REtrieval Conference (TREC2015)*.

Balabanović, Marko et Shoham, Yoav (1997). « Fab : content-based, collaborative recommendation ». *in* : *Communications of the ACM* **40**.*3*, p. 66–72.

Battistelli, Delphine et Teissèdre, Charles (2014). « Un outil d'observation du cheminement linguistique des événements médiatiques ». *in* : *Cahiers de praxématique 63*.

Billsus, Daniel et Pazzani, Michael J. (1999). « A hybrid user model for news story classification ». *in* : *UM99 User Modeling*. Springer, p. 99–108.

Binsztok, Henri et Gallinari, Patrick (2002). « Un algorithme en ligne pour la détection de nouveauté dans un flux de documents ». *in* : *6e Journées Internationales d'Analyse Statistique des Données Textuelles, JADT 2002*.

Bosredon, Bernard et Tamba, Irène (1992). « Thème et titre de presse : les formules bisegmentales articulées par un "deux points" ». *in* : *L'Information grammaticale* **54**.*1*, p. 36–44.

Bossard, Aurélien et Poibeau, Thierry (2008). « Regroupement automatique de documents en classes événementielles ». *in* : *15e Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2008)*, p. 201–210.

Boughanem, Mohand, Kraaij, Wessel et Nie, Jian-Yun (2004). « Modeles de langue pour la recherche d'information ». *in* : *Les systemes de recherche d'informations*, p. 163–182.

Brodt, Torben (2013). « The search for the best live recommender system ». *in* : *BARS'13 : Proceedings of the International SIGIR Workshop on Benchmarking Adaptive Retrieval and Recommender Systems,* p. 3–8.

Brodt, Torben et Hopfgartner, Frank (2014). « Shedding Light on a Living Lab : The CLEF NEWSREEL Open Recommendation Platform ». *in* : *IIiX'14 : Proceedings of the Information Interaction in Context Conference*. Regensburg, Germany : ACM, p. 223–226.

Brown, Ralf D., Pierce, Thomas, Yang, Yiming et Carbonell, Jaime G. (1999). « Link Detection – Results and Analysis ». *in* : *Topic Detection and Tracking Workshop – TDT3*.

Buckley, Chris, Dimmick, Darrin, Soboroff, Ian et Voorhees, Ellen M. (2007). « Bias and the limits of pooling for large collections ». *in* : *Information retrieval* **10**.*6*, p. 491–508.

Cantador, Iván, Bellogín, Alejandro et Castells, Pablo (2008). « News@hand : A semantic web approach to recommending news ». *in* : *International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems*. Springer, p. 279–283.

Charolles, Michel (2003). « De la topicalité des adverbiaux détachés en tête de phrase ». *in* : *Travaux de Linguistique : Revue Internationale de Linguistique Française 47*, p. 11–51.

Charon, Jean-Marie (2010). « De la presse imprimée à la presse numérique ». *in* : *Réseaux 2*, p. 255–281.

Chen, Francine, Farahat, Ayman et Brants, Thorsten (2004). « Multiple Similarity Measures and Source-Pair Information in Story Link Detection ». *in* : *Proceedings of Human Language Technology Conference (HLT-NAACL2044)*, p. 313–320.

Chen, Hsin-Hsi et Ku, Lun-Wei (2002). « An NLP & IR approach to topic detection ». *in* : *Topic detection and tracking*. Springer, p. 243–264.

Chy, Abu Nowshed, Ullah, Md Zia et Aono, Masaki (2015). « A Time and Context Aware Re-ranker for Microblog Retrieval ». *in* : 人工知能学会全国大会文集 **29**, p. 1–4.

Cieri, Chris, Graff, David, Liberman, Mark, Martey, Nii et Strassel, Stephanie (1999). « The TDT-2 text and speech corpus ». *in* : *Proceedings of the DARPA Broadcast News workshop*, p. 57–60.

Cleverdon, Cyril (1967). « The Cranfield tests on index language devices ». *in* : **19**.*6*, p. 173–194.

Clough, Paul et Sanderson, Mark (2013). « Evaluating the performance of information retrieval systems using test collections. » *in* : *Information Research* **18**.*2*.

Cohen, Jacob (1960). « A coefficient of agreement for nominal scale ». *in* : *Educ Psychol Meas* **20**, p. 37–46.

Conrad, Jack G et Bender, Michael (2016). « Semi-Supervised Events Clustering in News Retrieval. » *in* : *NewsIR@ ECIR*, p. 21–26.

Constant, Matthieu (2012). « Mettre les expressions multi-mots au coeur de l'analyse automatique de textes : sur l'exploitation de ressources symboliques externes ». Habilitation à Diriger des Recherches (HDR). Université Paris-Est.

Cooper, William S (1971). « A definition of relevance for information retrieval ». *in* : *Information storage and retrieval* **7**.*1*, p. 19–37.

Corney, David, Albakour, Dyaa, Martinez-Alvarez, Miguel et Moussa, Samir (2016). « What do a million news articles look like ? » *in* : *NewsIR@ ECIR*, p. 42–47.

Cotte, Pierre (1999). *Langage et linéarité*. Presses Univ. Septentrion.

Croft, W. Bruce, Cronen-Townsend, Stephen et Lavrenko, Victor (2001). « Relevance Feedback and Personalization : A Language Modeling Perspective ». *in* : *DELOS Workshop : Personalisation and Recommender Systems in Digital Libraries*. T. 3, p. 13.

Cselle, Gabor, Albrecht, Keno et Wattenhofer, Rogert (2007). « BuzzTrack : topic detection and tracking in email ». *in* : *Proceedings of the 12th international conference on Intelligent user interfaces*. ACM, p. 190–197.

Dagiral, Eric et Parasie, Sylvain (2010). « Vidéo à la une ! L'innovation dans les formats de la presse en ligne ». *in* : *Réseaux 2*, p. 101–132.

Daille, Béatrice (1994). « Approche mixte pour l'extraction de terminologie : statistique lexicale et filtres linguistiques ». Thèse de doctorat. Paris 7.

Dakka, Wisam, Gravano, Luis et Ipeirotis, Panagiotis (2012). « Answering general time-sensitive queries ». *in* : *IEEE Transactions on Knowledge and Data Engineering* **24**.*2*, p. 220–235.

De Gemmis, Marco, Lops, Pasquale, Musto, Cataldo, Narducci, Fedelucio et Semeraro, Giovanni (2015). « Semantics-aware content-based recommender systems ». *in* : *Recommender Systems Handbook*. Springer, p. 119–159.

Denis, Pascal et Sagot, Benoît (2010). « Exploitation d'une ressource lexicale pour la construction d'un étiqueteur morphosyntaxique état-de-l'art du français ». *in* : *Actes de TALN 2010*.

Dong, Anlei, Chang, Yi, Zheng, Zhaohui, Mishne, Gilad, Bai, Jing, Zhang, Ruiqiang, Buchner, Karolina, Liao, Ciya et Diaz, Fernando (2010). « Towards recency ranking in web search ». *in* : *Proceedings of the third ACM international conference on Web search and data mining*. ACM, p. 11–20.

Duda, Richard O., Hart, Peter E. et Stork, David G. (1973). *Pattern classification*. T. 2. Wiley New York.

Ehrmann, Maud (2008). « Les entités nommées, de la linguistique au TAL : statut théorique et méthodes de désambiguïsation ». Thèse de doctorat. Paris 7.

Fiscus, Jonathan G. et Doddington, George R. (2002). « Topic detection and tracking evaluation overview ». *in* : *Topic detection and tracking*. Springer, p. 17–31.

Fiscus, Jonathan G., Doddington, George R., Garofolo, John et Martin, Alvin (1999). « NIST's 1998 Topic Detection and Tracking evaluation (TDT2) ». *in* : *Proceedings of the 1999 DARPA Broadcast News Workshop*, p. 19–24.

Friburger, Nathalie, Maurel, Denis et Giacometti, Arnaud (2002). « Textual similarity based on proper names ». *in* : *Proceedings of the workshop Mathematical/Formal Methods in Information Retrieval*, p. 155–167.

Garg, Deepak et Sharma, Deepika (2012). « Information Retrieval on the Web and its Evaluation ». *in* : *Information Retrieval* **40**.*3*, p. 26–31.

Gomaa, Wael H. et Fahmy, Aly A. (2013). « A survey of text similarity approaches ». *in* : *International Journal of Computer Applications* **68**.*13*, p. 13–18.

Graff, David, Cieri, Chris, Strassel, Stephanie et Martey, Nii (1999). « The TDT-3 text and speech corpus ». *in* : *Proceedings of DARPA Broadcast News Workshop*, p. 57–60.

Gross, Gaston (1988). « Degré de figement des noms composés ». *in* : *Langages 90*, p. 57–72.

Gross, Maurice (1982). « Une classification des phrases «figées» du français ». *in* : *Revue québécoise de linguistique* **11**.*2*, p. 151–185.

Gross, Maurice et Senellart, Jean (1998). « Nouvelles bases statistiques pour les mots du français ». *in* : *4emes Journées internationales d'Analyse statistique des Données Textuelles (JADT'98)*, p. 335–349.

Gunawardana, Asela et Shani, Guy (2015). « Evaluating recommender systems ». *in* : *Recommender Systems Handbook*. Springer, p. 265–308.

Hatzivassiloglou, Vasileios, Gravano, Luis et Maganti, Ankineedu (2000). « An investigation of linguistic features and clustering algorithms for topical document clustering ». *in* : *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, p. 224–231.

Hatzivassiloglou, Vasileios, Klavans, Judith L. et Eskin, Eleazar (1999). « Detecting text similarity over short passages : Exploring linguistic feature combinations via machine learning ». *in* : *Proceedings of the 1999 joint sigdat conference on empirical methods in natural language processing and very large corpora*, p. 203–212.

Hill, Will, Stead, Larry, Rosenstein, Mark et Furnas, George (1995). « Recommending and evaluating choices in a virtual community of use ». *in* : *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM Press – Addison-Wesley Publishing Co., p. 194–201.

Hopfgartner, Frank, Brodt, Torben, Seiler, Jonas, Kille, Benjamin, Lommatzsch, Andreas, Larson, Martha, Turrin, Roberto et Serény, András (2015). « Benchmarking News Recommendations : The CLEF NewsREEL Use Case ». *in* : *SIGIR Forum* **49**.*2*, p. 129–136.

Hopfgartner, Frank, Kille, Benjamin, Lommatzsch, Andreas, Plumbaum, Till, Brodt, Torben et Heintz, Tobias (2014). « Benchmarking News Recommendations in a Living Lab ». *in* : *CLEF'14 : Proceedings of the 5th International Conference of the CLEF Initiative*. LNCS. Sheffield, UK : Springer Verlag, p. 250–267.

Kamba, Tomonari, Bharat, Krishna A. et Albers, Michael C. (1995). *The Krakatoa Chronicle – An interactive, personalized newspaper on the Web*. Rapport technique. Georgia Institute of Technology.

Kille, Benjamin, Hopfgartner, Frank, Brodt, Torben et Heintz, Tobias (2013). « The plista dataset ». *in* : *Proceedings of the 2013 International News Recommender Systems Workshop and Challenge*. ACM, p. 16–23.

KNIJNENBURG, Bart, MEESTERS, Lydia, MARROW, Paul et BOUWHUIS, Don (2009). « User-centric evaluation framework for multimedia recommender systems ». *in* : *International Conference on User Centric Media*. Springer, p. 366–369.

KNIJNENBURG, Bart, REIJMER, Niels J.M. et WILLEMSEN, Martijn C. (2011). « Each to His Own : How Different Users Call for Different Interaction Methods in Recommender Systems ». *in* : *Proceedings of the Fifth ACM Conference on Recommender Systems*. RecSys '11. Chicago, Illinois, USA : ACM, p. 141–148. ISBN : 978-1-4503-0683-6.

KNIJNENBURG, Bart, WILLEMSEN, Martijn C, GANTNER, Zeno, SONCU, Hakan et NEWELL, Chris (2012). « Explaining the user experience of recommender systems ». *in* : *User Modeling and User-Adapted Interaction* **22**.*4-5*, p. 441–504.

KOHAVI, Ron, LONGBOTHAM, Roger, SOMMERFIELD, Dan et HENNE, Randal M. (2009). « Controlled experiments on the web : survey and practical guide ». *in* : *Data mining and knowledge discovery* **18**.*1*, p. 140–181.

KONSTAN, Joseph A. et RIEDL, John (2012). « Recommender systems : from algorithms to user experience ». *in* : *User Modeling and User-Adapted Interaction* **22**.*1*, p. 101–123.

KRIEG-PLANQUE, Alice (2009). « À propos des "noms propres d'événement. Événementialité et discursivité ». *in* : *Les Carnets du Cediscor. Publication du Centre de recherches sur la didacticité des discours ordinaires 11*, p. 77–90.

KUMARAN, Giridhar et ALLAN, James (2004). « Text classification and named entities for new event detection ». *in* : *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, p. 297–304.

KUTUZOV, Andrey et KUZMENKO, Elizaveta (2016). « Cross-Lingual Trends Detection for Named Entities in News Texts with Dynamic Neural Embedding Models. » *in* : *NewsIR@ ECIR*, p. 27–32.

LANDIS, J. Richard et KOCH, Gary G. (1977). « The measurement of observer agreement for categorical data ». *in* : *Biometrics* **33**.*1*, p. 159–174.

LANDRAGIN, Frédéric (2004). « Saillance physique et saillance cognitive ». *in* : *Corela. Cognition, représentation, langage* **2**.*2*.

LANDRAGIN, Frédéric (2011). « De la saillance visuelle à la saillance linguistique ». *in* : *Saillance. Aspects linguistiques et communicatifs de la mise en évidence dans un texte*, p. 67–84.

LANDRAGIN, Frédéric (2012). « La saillance : questions méthodologiques autour d'une notion multifactorielle ». *in* : *Faits de langues* **39**.*1*, p. 15–31.

LANG, Ken (1995). « NewsWeeder : Learning to Filter Netnews ». *in* : *Proceedings of the 12th International Machine Learning Conference (ML95)*.

LEVY, Omer et GOLDBERG, Yoav (2014). « Neural word embedding as implicit matrix factorization ». *in* : *Advances in neural information processing systems*, p. 2177–2185.

Lin, Jimmy, Roegiest, Adam, Tan, Luchen, McCreadie, Richard, Voorhees, Ellen et Diaz, Fernando (2016). « Overview of the TREC 2016 real-time summarization track ». *in* : *Proceedings of the 25th Text REtrieval Conference, (TREC2016)*. T. 16.

Littré, Emile et Devic, L. Marcel (1869). *Dictionnaire de la langue française*. T. 4. L. Hachette et Cie.

Lops, Pasquale, De Gemmis, Marco et Semeraro, Giovanni (2011). « Content-based recommender systems : State of the art and trends ». *in* : *Recommender systems handbook*. Springer, p. 73–105.

Luhn, Hans Peter (1958). « The automatic creation of literature abstracts ». *in* : *IBM Journal of research and development* **2**.*2*, p. 159–165.

Ma, Hao, Liu, Xueqing et Shen, Zhihong (2016). « User Fatigue in Online News Recommendation ». *in* : *Proceedings of the 25th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, p. 1363–1372.

Manning, Christopher D., Raghavan, Prabhakar et Schütze, Hinrich (2008). « Evaluation in information retrieval ». *in* : *Introduction to information retrieval*, p. 151–175.

Martin, Alvin, Doddington, George R., Kamm, Terri, Ordowski, Mark et Przybocki, Mark (1997). *The DET curve in assessment of detection task performance*. Rapport technique. DTIC Document.

Martinez, Miguel, Kruschwitz, Udo, Kazai, Gabriella, Hopfgartner, Frank, Corney, David, Campos, Ricardo et Albakour, Dyaa (2016). « Report on the 1st International Workshop on Recent Trends in News Information Retrieval (NewsIR16) ». *in* : *SIGIR Forum* **50**.*1*, p. 58–67. issn : 0163-5840.

Mathet, Yann et Widlöcher, Antoine (2016). « Évaluation des annotations : ses principes et ses pièges ». *in* : *TAL et Ethique* **57**.*2*, p. 73–98.

Mathieu-Colas, Michel (1996). « Essai de typologie des noms composés français ». *in* : *Cahiers de lexicologie 69*, p. 71–125.

McNee, Sean M, Riedl, John et Konstan, Joseph A (2006). « Making recommendations better : an analytic model for human-recommender interaction ». *in* : *CHI'06 extended abstracts on Human factors in computing systems*. ACM, p. 1103–1108.

Mikolov, Tomas, Sutskever, Ilya, Chen, Kai, Corrado, Greg S. et Dean, Jeff (2013). « Distributed representations of words and phrases and their compositionality ». *in* : *Proceedings of the 26th International Conference on Neural Information Processing Systems (NIPS'13)*, p. 3111–3119.

Miller, George A. (1995). « WordNet : a lexical database for English ». *in* : *Communications of the ACM* **38**.*11*, p. 39–41.

Moirand, Sophie (2007). *Les discours de la presse quotidienne (observer, analyser, comprendre)*. Presses universitaires de France (Linguistique nouvelle).

Mooers, Calvin N. (1948). « Application of random codes to the gathering of statistical information ». Thèse de doctorat. Massachusetts Institute of Technology.

Moreau, Fabienne et Claveau, Vincent (2006). « Extension de requêtes par relations morphologiques acquises automatiquement ». *in* : *COnférence en Recherche d'Informations et Applications - CORIA 2006*, p. 181–192.

Moschitti, Alessandro et Basili, Roberto (2004). « Complex linguistic features for text classification : A comprehensive study ». *in* : *Advances in Information Retrieval : 26th European Conference on IR Research, ECIR 2004*. Springer, p. 181–196.

Nazarenko, Adeline, Zargayouna, Haïfa, Hamon, Olivier et Van Puymbrouck, Jonathan (2009). « Évaluation des outils terminologiques : enjeux, difficultés et propositions ». *in* : *Traitement Automatique des Langues, ATALA* **50**.*1 varia*, p. 257–281.

Névéol, Aurélie, Zeng, Kelly et Bodenreider, Olivier (2006). « Besides precision & recall : exploring alternative approaches to evaluating an automatic indexing tool for MEDLINE. » *in* : *Annual Symposium Proceedings, American Medical Informatics Association (AMIA 2006)*.

Palmer, David D. (2000). « Tokenisation and sentence segmentation ». *in* : *Handbook of Natural Language Processing*. CRC Press, p. 11–35.

Paternostre, Marjorie, Francq, Pascal, LAMORAL, Julien, Wartel, David et Saerens, Marco (2002). *Carry, un algorithme de désuffixation pour le français*. Rapport technique du projet Galilei.

Pazzani, Michael J., Muramatsu, Jack et Billsus, Daniel (1996). « Syskill & Webert : Identifying interesting web sites ». *in* : *Proceedings of th 13th National Conference on Artificial Intelligence*, p. 54–61.

Peirce, Charles S. (1931). *The Collected Papers of Charles Sanders Peirce, Vol. I : The Principles of Philosophy*. Sous la dir. de Charles Hartshorne et Paul Weiss. Cambridge : Harvard University Press, p. 423.

Phelan, Owen, McCarthy, Kevin et Smyth, Barry (2009). « Using twitter to recommend real-time topical news ». *in* : *Proceedings of the third ACM conference on Recommender systems*. ACM, p. 385–388.

Piworwarski, Benjamin (2003). « Techniques d'apprentissage pour le traitement d'informations structurées : application à la recherche d'information ». Thèse de doctorat. Paris 6.

Porter, Martin F. (1980). « An algorithm for suffix stripping ». *in* : *Program* **14**.*3*, p. 130–137.

Pu, Pearl, Chen, Li et Hu, Rong (2011). « A user-centric evaluation framework for recommender systems ». *in* : *Proceedings of the fifth ACM conference on Recommender systems*. ACM, p. 157–164.

Pu, Pearl, Chen, Li et Hu, Rong (2012). « Evaluating recommender systems from the user's perspective : survey of the state of the art ». *in* : *User Modeling and User-Adapted Interaction* **22**.*4*, p. 317–355.

Qiu, Jing, Liao, Lejian et Li, Peng (2009). « News recommender system based on topic detection and tracking ». *in* : *Rough Sets and Knowledge Technology : 4th International Conference, RSKT 2009*. Springer, p. 690–697.

Resnick, Paul, Iacovou, Neophytos, Suchak, Mitesh, Bergstrom, Peter et Riedl, John (1994). « GroupLens : an open architecture for collaborative filtering of netnews ». *in* : *Proceedings of the 1994 ACM conference on Computer supported cooperative work*. ACM, p. 175–186.

Ringoot, Roselyne (2014). *Analyser le discours de presse*. Armand Colin.

Robertson, Stephen E. (1977). « The probability ranking principle in IR ». *in* : *Journal of documentation* **33**.*4*, p. 294–304.

Robertson, Stephen E. et Walker, Steve (1994). « Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval ». *in* : *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*. Springer-Verlag New York, Inc., p. 232–241.

Salem, André (1986). « Segments répétés et analyse statistique des données textuelles ». *in* : *Histoire & Mesure*, p. 5–28.

Salton, Gerard (1968). *Automatic information organization and retrieval*. McGraw-Hill computer science series. McGraw-Hill.

Salton, Gerard (1971). *The SMART Retrieval System - Experiments in Automatic Document Processing*. Prentice Hall.

Salton, Gerard, Wong, Anita et Yang, Chung-Shu (1975). « A vector space model for automatic indexing ». *in* : *Communications of the ACM* **18**.*11*, p. 613–620.

Salton, Gerard et Yang, Chung-Shu (1973). « On the specification of term values in automatic indexing ». *in* : *Journal of documentation* **29**.*4*, p. 351–372.

Sanderson, Mark et Van Rijsbergen, CJ (1991). « Nrt-news retrieval tool ». *in* : *Electronic Publishing : Origination, Dissemination, and Design* **4**.*4*, p. 205–217.

Saracevic, Tefko (1975). « Relevance : A review of and a framework for the thinking on the notion in information science ». *in* : *Journal of the American Society for information science* **26**.*6*, p. 321–343.

Sarwar, Badrul, Karypis, George, Konstan, Joseph et Riedl, John (2001). « Item-based collaborative filtering recommendation algorithms ». *in* : *Proceedings of the 10th international conference on World Wide Web*. ACM, p. 285–295.

Schamber, Linda (1994). « Relevance and information behavior. » *in* : *Annual review of information science and technology (ARIST)* **29**, p. 3–48.

Schmid, Helmut (1994). « Probabilistic Part-of-Speech Tagging Using Decision Trees ». *in* : *International Conference on New Methods in Language Processing*. Manchester, UK, p. 44–49.

Shardanand, Upendra et Maes, Pattie (1995). « Social Information Filtering : Algorithms for Automating "Word of Mouth" ». *in* : *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '95, p. 210–217.

Sparck Jones, Karen (1972). « A statistical interpretation of term specificity and its application in retrieval ». *in* : *Journal of documentation* **28**.*1*, p. 11–21.

Sparck Jones, Karen (1973). « Index term weighting ». *in* : *Information storage and retrieval* **9**.*11*, p. 619–633.

Sparck Jones, Karen et Van Rijsbergen, Cornelis Joost (1975). « Report on the need for and provision of an "ideal" information retrieval test collection ». *in* : *Computer Laboratory*.

Sparck Jones, Karen et Van Rijsbergen, Cornelis Joost (1976). « Information retrieval test collections ». *in* : *Journal of documentation* **32**.*1*, p. 59–75.

Sparck Jones, Karen, Walker, Steve et Robertson, Stephen E. (2000). « A probabilistic model of information retrieval : development and comparative experiments : Part 2 ». *in* : *Information processing & management* **36**.*6*, p. 809–840.

Steimberg, Laura Calabrese (2006). « La construction de la mémoire historico-médiatique à travers les désignations d'événements ». *in* : *Studies Van de BKL 2006*.

Steimberg, Laura Calabrese (2012). « L'acte de nommer : nouvelles perspectives pour le discours médiatique ». *in* : *Langage et société 2*, p. 29–40.

Tavakolifard, Mozhgan et al. (2013). « Workshop and Challenge on News Recommender Systems ». *in* : *Proceedings of the 7th ACM Conference on Recommender Systems*. RecSys '13. Hong Kong, China : ACM, p. 481–482. isbn : 978-1-4503-2409-0.

Teltzrow, Maximilian et Kobsa, Alfred (2004). « Impacts of User Privacy Preferences on Personalized Systems ». *in* : *Designing Personalized User Experiences in eCommerce*. Sous la dir. de Clare-Marie Karat, Jan O. Blom et John Karat. Dordrecht : Springer Netherlands, p. 315–332.

Thibaut, Jean-Pierre (1997). « Similarité et catégorisation ». *in* : *L'année psychologique* **97**.*4*, p. 701–736.

Tsagkias, Manos, De Rijke, Maarten et Weerkamp, Wouter (2011). « Linking online news and social media ». *in* : *Proceedings of the fourth ACM international conference on Web search and data mining*. ACM, p. 565–574.

Tversky, Amos (1977). « Features of similarity. » *in* : *Psychological review* **84**.*4*, p. 327.

Voorhees, Ellen M. (2000). « Variations in relevance judgments and the measurement of retrieval effectiveness ». *in* : *Information processing & management* **36**.*5*, p. 697–716.

VOORHEES, Ellen M. (2001). « The philosophy of information retrieval evaluation ». *in* : *Evaluation of Cross-Language Information Retrieval Systems : Second Workshop of the Cross-Language Evaluation Forum, CLEF 2001*. Springer, p. 355–370.

WAYNE, Charles L. (2000). « Multilingual Topic Detection and Tracking : Successful Research Enabled by Corpora and Evaluation ». *in* : *Proceedings of the LREC-00*.

WILLEMSEN, Martijn C, KNIJNENBURG, Bart, GRAUS, Mark P, VELTER-BREMMERS, LC et FU, Kai (2011). « Using latent features diversification to reduce choice difficulty in recommendation lists ». *in* : *RecSys* **11**, p. 14–20.

WILSON, Patrick (1973). « Situational relevance ». *in* : *Information storage and retrieval* **9**.*8*, p. 457–471.

YANG, Yiming, CARBONELL, Jaime G., BROWN, Ralf D., PIERCE, Thomas, ARCHIBALD, Brian T. et LIU, Xin (1999). « Learning Approaches for Detecting and Tracking News Events ». *in* : *IEEE Intelligent Systems*, p. 32–43.

YANG, Yiming, CARBONELL, Jaime, BROWN, Ralf D., LAFFERTY, John, PIERCE, Thomas et AULT, Thomas (2002). « Multi-strategy learning for topic detection and tracking ». *in* : *Topic detection and tracking*. Springer, p. 85–114.

YU, Huan, WANG, Ying, FAN, Yaning, MENG, Sachula et HUANG, Rui (2017). « Accuracy Is Not Enough : Serendipity Should Be Considered More ». *in* : *International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing*. Springer, p. 231–241.

ZHAI, Chengxiang et LAFFERTY, John (2001). « A study of smoothing methods for language models applied to ad hoc information retrieval ». *in* : *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, p. 334–342.

ZHENG, Zhaohui, SRIHARI, Rohini et SRIHARI, Sargur (2003). « A feature selection framework for text filtering ». *in* : *Data Mining, 2003. ICDM 2003. Third IEEE International Conference on*. IEEE, p. 705–708.

ZIPF, George (1935). *The Psychobiology of Language : An Introduction to Dynamic Philology*. Cambridge, Mass. : M.I.T. Press.