Hypergraphes multimédias dirigés navigables : construction et exploitation

Rémi Bois

ANNÉE 2017

UNIVERSITÉ DE RENNES 1

UNIVERSITE BRETAGNE LOIRE

THÈSE / UNIVERSITÉ DE RENNES 1
sous le sceau de l'Université Bretagne Loire

pour le grade de

DOCTEUR DE L'UNIVERSITÉ DE RENNES 1

Mention : Informatique

École doctorale Matisse

présentée par

Rémi Bois

préparée à l'unité de recherche 6074 IRISA
Institut de recherche en informatique et systèmes aléatoires
Université de Rennes 1

Hypergraphes multimédias dirigés navigables : construction et exploitation

Thèse soutenue à Rennes
le 21 Décembre 2017

devant le jury composé de :

Bénédicte LE GRAND
Professeur, Univ. Panthéon Sorbonne, CRI / présidente

Patrice BELLOT
Professeur, Univ. d'Aix-Marseille, CNRS, LSIS / rapporteur

Xavier TANNIER
Professeur, Univ. Paris-Sud, CNRS, LIMSI / rapporteur

Jean CARRIVE
INA / examinateur

Emmanuel MORIN
Professeur, Univ. Nantes, LS2N / encadrant

Guillaume GRAVIER
Directeur de recherche, CNRS, IRISA & INRIA Rennes / directeur de thèse

Pascale SÉBILLOT
Professeur, INSA Rennes, IRISA & INRIA Rennes / directrice de thèse

Remerciements

Ces quelques lignes sont pour moi l'occasion de coucher par écrit des remerciements rarement formulés, et pourtant amplement mérités. D'abord pour Pascale, Guillaume et Emmanuel, qui m'ont guidé tout au long de ces trois années. Merci pour votre confiance, votre exigence et votre bienveillance qui m'ont permis de sans cesse progresser dans les meilleures conditions possibles. Ensuite pour les membres de mon jury de thèse pour avoir accepté d'évaluer mon travail.

Cette thèse a été réalisée au sein du projet LIMAH[1]. J'ai une pensée pour chacune des personnes y ayant participé, nos discussions ont toujours été passionnantes et suivre l'avancée de vos travaux a été un plaisir. L'équipe LinkMedia qui m'a accueilli à Rennes pendant ces trois ans a été la source de nombreux conseils et d'une ambiance de travail agréable quotidienne. Merci à tous les permanents, à Aurélie dont l'aide a été inestimable, et à tous les doctorants, post-doc et ingénieurs que j'ai eu la chance de côtoyer au sein de l'équipe. Je remercie spécialement ceux de l'équipe avec qui j'ai partagé des travaux de recherche, à savoir Anca, Vedran, Mikail, Mateus, Ahmet, Ronan et Arnaud. Merci également aux collègues avec qui j'ai eu la chance d'enseigner, et particulièrement Delphine et Thomas. Votre dévouement pour vos jeunes étudiants est source d'inspiration. Enfin, je clos ces remerciements dédiés à l'équipe par ceux avec qui j'ai partagé d'innombrables cafés : le maître décorateur Cédric qui, j'en suis sûr, parviendra à finir de recouvrir les murs de notre bureau avant la fin de sa thèse, et le maître joueur Clément, qui a eu l'infortune de m'avoir en partenaire de jeu, menant de nombreuses fois à sa mort virtuelle.

J'ai eu l'occasion pendant ces trois années de rencontrer de nombreux doctorants par le biais de l'association Nicomaque. J'en remercie tous les membres, et spécialement Victorien, Roselyne, Mathilde, Dominique, Lida et Yann avec qui j'ai eu l'occasion de mener des projets passionnants. Lisa, un merci particulier à toi, pour les mêmes raisons et pour m'avoir présenté celle avec qui je coule des jours heureux.

J'ai décidé dans m'engager dans cette thèse après des études captivantes menées à Nantes. Durant mon master là-bas, j'ai eu l'opportunité de rencontrer ceux qui m'ont transmis leur passion pour la recherche. Merci à l'ensemble de l'équipe TALN pour les précieux enseignements qu'ils ont su prodiguer. Je pense tout particulièrement à Florian, dont les conseils inestimables m'ont donné envie de consacrer trois années supplémentaires de ma vie à faire de la recherche. Je pense également aux autres étudiants du master, à savoir Grégoire, Agathe, Soufian, Joseph, Hugo, Noémi et Loïc avec qui j'ai partagé des moments mémorables. Nathalie, si tu n'as partagé nos salles de classes que quelques mois, tu es malgré tout un membre irremplaçable de notre petite équipe. Ces deux années en votre compagnie ont été à la fois stimulantes et remplies de bonheur.

Un autre facteur de réussite de cette thèse réside en la présence continuelle d'amis précieux. Un grand merci à Benjamin, Romain, Guillaume et Corentin pour les bons moments passés ensemble depuis plus de 10 ans. Votre amitié m'est précieuse. Merci également à tous ceux que j'ai rencontré depuis le début de mes études, et avec qui je conserve des liens très forts. Marie-Charlotte, Chris, Quentin, Eric, Carl, Clément, Romain, Kevin, Nicolas, Gwen, Phil, Baptiste, c'est toujours un immense plaisir de vous retrouver.

Il est temps de conclure ces remerciements avec l'expression de mon amour pour mes parents qui n'ont eu de cesse de m'encourager et de me soutenir. Maman, papa, merci. Merci également à ma jumelle Laura, dont j'admire la patience, la ténacité, et la gentillesse. Enfin, merci à toi, Caroline, pour être présente chaque jour à mes côtés. Je t'aime.

1. Ce travail a bénéficié d'une aide de l'État attribuée au labex COMIN LABS et gérée par l'Agence Nationale de la Recherche au titre du programme « Investissements d'avenir » portant la référence ANR-10-LABX-07-01.

Table des matières

Introduction générale

L'accès à une information diversifiée et de qualité est un enjeu essentiel pour l'ensemble de la société. Depuis plusieurs années, les sources d'informations se multiplient, entraînant de fait une multiplication des points de vue, qu'ils soient rapportés par des médias établis au travers de journaux, par d'apprentis journalistes via des blogs, ou par des citoyens utilisant les réseaux sociaux comme seul mode de diffusion. Les moyens de consommation de cette actualité se sont également diversifiés, et de récentes études rapportent que la part de la population s'informant via les réseaux sociaux est en nette augmentation (Gottfried et Shearer, 2016). Cette multiplicité a néanmoins certains effets néfastes, tels que la large diffusion de fausses informations (*fake news*), la difficulté ressentie par le grand public à s'orienter dans la masse d'informations disponibles, ou encore les efforts nécessaires aux professionnels des médias (journalistes, attachés de presse, ...) pour trouver des éléments d'information précis. Dans cette thèse nous nous intéressons aux deux dernières problématiques, qui ne trouvent pas à l'heure actuelle de réponse satisfaisante.

Peu d'outils existent aujourd'hui pour permettre au grand public d'explorer efficacement les nombreux documents d'actualités publiés chaque jour. Face à des volumes de publication gigantesques, multisources et multimédias, trois approches principales coexistent. La première consiste à utiliser un média de référence, qui se charge lui-même de hiérarchiser et de sélectionner les informations qu'il juge pertinentes. Ce média peut correspondre à une entité de presse écrite (Le Monde, Le Figaro, Libération, ...), télévisuelle (journaux télévisés, émissions d'actualités, ...) ou radiophonique (bulletins d'information, chroniques d'information, ...). On peut alors parler d'approche verticale, dans laquelle l'information est générée, mise à disposition, et triée par ces entités avant d'être consommée par le grand public. La deuxième approche consiste en une approche horizontale, dans laquelle c'est le grand public qui hiérarchise, sélectionne, voire génère l'information. C'est le modèle des sites communautaires comme AgoraVox ou Reddit, et des réseaux sociaux comme Facebook ou Twitter. La troisième et dernière approche consiste à offrir un large spectre des publications des différents médias et à laisser l'utilisateur sélectionner ses sujets d'intérêt ainsi que ses sources préférées. C'est le modèle des agrégateurs, qui rassemblent les articles de presse discutant d'un même événement au sein de groupes distincts et laisse à l'utilisateur la liberté de choisir quelle source consulter au sein de chaque groupe. Ce dernier modèle, rendu possible par l'utilisation d'interfaces web efficaces, permet notamment de répondre à l'envie exprimée par les citoyens européens d'avoir accès à plusieurs points de vue (Newman et al., 2016). Dans cette thèse, nous proposons d'étendre la notion d'agrégateurs en construisant une structuration plus riche que le regroupement d'informations au sein de groupes distincts. Nous y déve-

loppons l'idée d'une approche par graphe, dans laquelle les documents d'information, multimédias et multisources, sont reliés de façon à faciliter une exploration éclairée de l'actualité.

Les besoins des professionnels de l'information sont variables. On peut rapprocher les besoins des attachés de presse de ceux du grand public, en cela qu'ils ont régulièrement des besoins d'exploration de l'actualité afin de constituer des panoramas représentatifs de l'opinion publique. Les journalistes sont, au contraire, généralement en recherche d'informations précises et ciblées, afin d'enrichir leurs articles en les contextualisant. L'outil privilégié pour ces professionnels consiste en l'utilisation de moteurs de recherche, souvent construits sur des archives de presse. Ce constat est à nuancer en fonction des journalistes, le journalisme d'investigation faisant largement appel à des processus d'exploration de l'information. Cela a notamment été le cas avec l'affaire des Panama Papers, pour laquelle des outils d'exploration fondés sur des graphes [1] ont été utilisés.

Cette thèse se déroule dans le cadre du projet de recherche CominLabs *LInking Media in Acceptable Hypergraphs* (LIMAH), qui vise à enrichir la consultation de grandes collections multimédias selon plusieurs angles, notamment celui de l'exploration d'actualités et la structuration de cours en ligne. Ce projet regroupe des équipes scientifiques issues de plusieurs domaines de recherche. L'IRISA et le LS2N abordent notamment les problématiques multimédias et langagières, ainsi que l'étude de sentiments appliquée aux réactions du grand public à l'information. L'IODE s'intéresse aux implications légales de la structuration de collections et notamment à la place du lien hypertexte dans la législation européenne. Le PREFICs et le CRPCC s'intéressent quant à eux à la perception de l'utilité d'outils d'exploration d'actualités par des utilisateurs, ainsi qu'à la modification éventuelle des pratiques professionnelles liées à l'utilisation de ces nouveaux outils. En ce qui concerne les contributions de cette thèse, c'est-à-dire l'exploration de collections journalistiques, les domaines de recherche suivants sont abordés : d'une part les sciences sociales, à travers l'étude de la production journalistique, et d'autre part le traitement des langues et la recherche d'information, qui se sont notamment intéressés au regroupement d'articles de presse similaires tels que proposés par les agrégateurs, à l'ordonnancement temporel d'articles discutant des événements successifs, voire à la description de relations de causalité entre événements. Si l'ordonnancement temporel est abordé dans le cadre de cette thèse, l'approche que nous proposons diffère par la structuration en graphe des collections d'actualité. Un autre domaine pertinent est le multimédia et notamment la recherche d'information multimédia. Cette discipline s'intéresse notamment aux documents multimédias d'actualités, mais travaille essentiellement avec des données monosources, issues d'un unique organe de presse. Une branche de la recherche d'information multimédia, appelée hyperliage (ou *hyperlinking*) constitue la base sur laquelle cette thèse est construite. L'hyperliage consiste à construire automatiquement des liens entre documents multimédias. Nous étendons ce concept en l'appliquant à l'entièreté d'une collection afin d'obtenir un hypergraphe, et nous intéressons notamment à ses caractéristiques topologiques. Enfin, le domaine de recherche de l'ergonomie et des interfaces utilisateurs s'intéresse à l'utilisabilité des systèmes informatiques. Dans le cadre de cette thèse, nous conduisons, en collaboration étroite avec le CRPCC, des études utilisateurs permettant de vérifier la pertinence des systèmes que nous construisons.

Nous proposons dans cette thèse des améliorations de l'état de l'art selon trois axes principaux : une structuration de collections d'actualités à l'aide de graphes multisources

1. www.data.blog.lemonde.fr/2016/04/08/panama-papers-un-defi-technique-pour-le-journalisme-de-donnees/

et multimodaux fondée sur la création de liens inter-documents, son association à une diversité importante des liens, et enfin l'ajout d'un typage des liens créés permettant d'expliciter la relation existant entre deux documents.

Ce manuscrit est composé de trois parties organisées comme suit :

La première partie expose les enjeux et outils disponibles pour l'exploration d'actualités. Nous décrivons dans le premier chapitre les populations concernées par l'exploration d'actualités ainsi que les outils déjà existants. Nous détaillons dans le deuxième chapitre les solutions académiques au travers d'un état de l'art. Enfin, le projet LIMAH, dans lequel cette thèse s'inscrit, est présenté dans le chapitre 3, avec une attention particulière portée sur la construction du corpus qui fût la première étape de cette thèse.

La seconde partie développe la construction d'hypergraphes pour l'exploration d'actualités. La notion d'hypergraphe y est précisée au sein du chapitre 4 au travers d'un comparatif avec les paradigmes de moteur de recherche et de recommandation. Le critère de navigabilité d'un graphe est également présenté dans ce chapitre avant d'introduire la notion d'explorabilité. Nous exposons ensuite dans le chapitre 5 une nouvelle méthode pour la construction de graphes d'actualités et la comparons à deux méthodes standards, K-NN et $\mathcal{E}$-NN. Enfin, nous étudions dans le chapitre 6 les différents moyens à disposition pour améliorer la diversité des liens entre documents vidéos, au travers de modèles combinant deux modalités. Nous y proposons également une troisième approche permettant un meilleur contrôle sur la diversité des liens créés ainsi qu'une étude visant à faciliter l'évaluation de la diversité obtenue.

La troisième partie s'intéresse à l'enrichissement des graphes d'actualités par le typage des liens. Celui-ci vise à éclairer l'utilisateur lors de sa navigation en explicitant la relation existant entre deux documents. Dans le chapitre 7, nous décrivons la typologie retenue pour notre cas d'étude, et discutons des différentes méthodes permettant d'automatiser ce typage. Dans le chapitre 8, nous validons l'intérêt du typage et de l'utilisation d'hypergraphes plutôt que de moteurs de recherche via des études utilisateurs impliquant journalistes et étudiants journalistes.

Première partie

Enjeux et moyens pour l'exploration d'actualités

Chapitre 1

Explorer l'actualité : un enjeu pour les professionnels et le grand public

Introduction

La consultation d'actualités par le grand public et les professionnels est plus que jamais réalisée par de nombreux canaux. Le grand public s'informe depuis longtemps via son poste de télévision, les journaux papiers, la presse en ligne. Ces sources, classiques, peuvent le plus souvent être consultées en ligne, au travers des sites des organes de presse, de *replays*, ou de *podcasts*. Les sources textuelles sont pour la plupart rassemblées par des agrégateurs d'actualités, qui visent à proposer une vue d'ensemble de l'actualité. À ces canaux se sont récemment ajoutés les réseaux sociaux et les sites communautaires, deux nouveaux médias qui contribuent largement à la diffusion et la mise en avant d'actualités. Ainsi, un article publié sur un blog peu connu peut avoir un impact plus important sur la société qu'un article d'un média établi s'il est suffisamment partagé sur les réseaux sociaux. Si ces différents canaux peuvent être considérés comme complémentaires, force est de constater qu'il n'existe pas aujourd'hui d'outils performants permettant simultanément de bénéficier de la grande quantité de documents consultables et de donner du sens aux informations.

Les professionnels de l'information, tels que les journalistes, *community managers*, ou attachés de presse, ont une activité professionnelle nécessitant une consultation rapide et efficace d'informations. Cette catégorie a des besoins plus importants que le grand public et dispose généralement de grandes quantités de données, sous forme d'archives journalistique, le plus souvent consultées par le biais de moteurs de recherche. Paradoxalement, ces professionnels ne disposent pas ou peu d'outils plus performants que le grand public pour mener à bien leurs tâches et utilisent sensiblement les mêmes types de services.

Dans ce chapitre, nous commençons par expliquer comment deux catégories de personnes sont concernées par la consultation d'actualités : le grand public et les professionnels de l'information. Après avoir présenté ces deux catégories, leurs pratiques et leurs besoins, nous décrivons plus finement les outils dont ils disposent pour parcourir et traiter l'actualité. Si l'objectif n'est pas de dresser ici un inventaire exhaustif des outils disponibles, nous en feront une liste partielle et une analyse rapide afin de mettre en avant les paradigmes qu'ils mettent en jeu et leurs faiblesses. Nous présentons ensuite

une étude que nous avons menée en collaboration avec le CRPCC et qui vise à évaluer les attentes des professionnels de l'information en terme d'outillage et à éclairer leurs manques (Gravier et al., 2016).

1.1 Populations concernées

Nous décrivons ici deux populations particulièrement concernées par l'exploration d'actualités : le grand public et les professionnels de l'information.

1.1.1 Grand public

Le grand public est friand d'actualités. Les sources à sa disposition se sont diversifiées, et si 74 % des Français utilisent leur poste de télévision comme outil d'accès à l'information, ils sont également 71 % à utiliser Internet pour s'informer. La différence est encore plus importante chez les moins de 35 ans qui ne sont plus que 61 % à consulter les informations à la télévision. La tendance est identique dans les 26 pays étudiés par Newman et al. (2016), et montre que chez les moins de 45 ans, la principale source d'information est le web. D'autre part, alors qu'Internet était perçu comme une source peu crédible il y a quelques années, sa crédibilité auprès du grand public européen augmente progressivement et s'établissait à 36 % en 2014, en progression de 2 points par rapport à l'année précédente (TNS, 2015). Ce taux de confiance important est néanmoins à relativiser en France, puisqu'il ne s'y établit qu'à 27 %. Ce manque de confiance se retrouve également dans les autres médias en France (34 % de confiance envers la télévision contre 50 % en Europe, 55 % de confiance envers la radio contre 58 % en Europe). Seule la presse écrite classique est plébiscitée par les Français qui la jugent crédible à 45 % contre 43 % en Europe. La table 1.1 décrit la consommation médiatique en fonction de la catégorie d'âge en France telle que rapportée dans (Newman et al., 2016).

	18-24	25-34	35-44	45-54	55+
Web (réseaux sociaux inclus)	64 %	57 %	47 %	36 %	25 %
Réseaux sociaux	28 %	17 %	12 %	8 %	5 %
Radio	5 %	6 %	7 %	8 %	8 %
Presse écrite	6 %	6 %	7 %	8 %	12 %
Télévision	24 %	29 %	37 %	46 %	53 %

TABLE 1.1 – Principales sources d'information selon la catégorie d'âge en France.

Si les consommateurs d'actualités en ligne consultent essentiellement des articles en provenance de quelques médias de confiance (dont la liste peut varier d'utilisateur en utilisateur), ils sont confrontés à des quantités de données difficiles à appréhender (Sturgill et al., 2010). La plupart d'entre eux utilisent donc comme point de départ un agrégateur d'actualités ou un moteur de recherche (Newman et al., 2015), qui permettent de limiter, ou à défaut d'ordonner, les informations selon leur pertinence (Holton et Chyi, 2012). Ainsi, Google News [1] est utilisé de façon hebdomadaire par 11 % des personnes interrogées en France, Yahoo News [2] par 8 % et MSN Actualités [3] par 7 % d'entre elles.

1. www.news.google.com
2. www.yahoo.com/news/
3. www.msn.com/fr-fr/actualite

Seuls deux médias (Le Monde et 20 Minutes) obtiennent des scores de fréquentation supérieurs à l'agrégateur de Google (Newman et al., 2015). L'arrivée de ces agrégateurs comme intermédiaire pour accéder à l'information a d'abord été perçue négativement par les différents médias qui craignaient que leurs lecteurs n'atteignent jamais leurs sites, et se contentent plutôt des titres et descriptions succinctes directement disponibles sur les agrégateurs. De récentes études semblent montrer que, malgré le fait que la plupart des lecteurs se contentent effectivement de lire les grands titres, les agrégateurs ont en réalité un effet positif sur la fréquentation des sites d'actualité. En effet, suite à la fermeture de Google News en Espagne, une chute de 11 % de la fréquentation des sites d'actualité a été enregistrée selon Calzada et Gil (2016). Cet effet positif est particulièrement marqué pour certains types de contenu (*e.g.* actualités locales (Athey et Mobius, 2012)) et pour le public étranger (*e.g.* expatriés (Calzada et Gil, 2016)). Selon cette même étude, il n'y aurait pas d'effet de substitution des agrégateurs aux sites d'actualité, et donc, les agrégateurs n'auraient que des effets positifs pour les médias en ligne.

Si les agrégateurs permettent la multiplicité des sources en regroupant la plupart des médias traditionnels en un même endroit, ils ignorent en général les articles écrits par de simples citoyens non rattachés à un groupe de presse. Ce journalisme citoyen (Kaufhold et al., 2010), écrit en dehors des salles de rédaction, fait généralement l'objet d'une publication au travers de blogs (Pledel, 2006) et vise à apporter une vision souvent différente des médias traditionnels (Wall, 2005). Si cette pratique a d'abord été perçue négativement par les journalistes professionnels, un consensus existe désormais sur leur complémentarité (Lasica, 2003; Lowrey, 2006) et de nombreux journaux traditionnels hébergent eux-mêmes des blogs. Certaines plateformes, telles AgoraVox [4], visent à étendre ce phénomène en offrant une plateforme de publication participative et citoyenne, les choix éditoriaux étant réalisés par les auteurs d'articles eux-mêmes, et la rédaction d'articles étant ouverte à tous. Cette modération commune rapproche ces plateformes des sites communautaires, qui sont discutés plus loin.

Depuis quelques années, les réseaux sociaux occupent une place de plus en plus importante en tant que source d'information pour le grand public (Bakshy et al., 2012; Kwak et al., 2010), notamment les plus jeunes (TNS, 2015). Ils sont considérés comme utiles pour la consultation d'actualités politiques, 52 % des Européens (et 48 % des Français, en progression de 6 points sur un an) estimant que les réseaux sociaux sont « un moyen moderne de rester au courant des affaires politiques ». Ils sont également logiquement plébiscités pour le partage d'opinion, 51 % des Européens (et 53% des Français) pensant qu'ils sont « un bon moyen de dire ce qu'on pense des questions politiques ». Ils sont néanmoins perçus comment peu fiables par 44% des Européens (et 57 % des Français). Sur ces réseaux, aucune différence n'est faite entre sources traditionnelles et journalisme citoyen, ni même entre information et divertissement (y compris la parodie), menant à de fortes critiques sur leur fiabilité (Allcott et Gentzkow, 2017). Néanmoins, ils sont fréquemment considérés comme un moyen efficace d'inciter le grand public à s'intéresser à l'actualité et à participer à la vie politique (Gil de Zúñiga et al., 2012).

À mi-chemin entre les réseaux sociaux et les médias traditionnels, on trouve les sites communautaires, au sein desquels les utilisateurs partagent des contenus, les filtrent, les mettent en avant, les commentent, les débattent. Si la plupart de ces sites ou des communautés qu'ils hébergent s'intéressent à des thématiques particulières comme l'arrivée de nouveaux produits (ProductHunt) ou la programmation informatique (HackerNews), d'autres accordent une importance particulière à l'actualité. C'est notamment le cas de

4. www.agoravox.fr

Reddit, 8e site le plus visité au monde selon Alexa [5], 4e site le plus visité aux États-Unis, et 7e site le plus visité en France [6].

Ces différents moyens largement utilisés par le grand public afin d'explorer l'actualité sont décrits plus finement section 1.2.

1.1.2 Professionnels de l'information

De nombreuses professions nécessitent une utilisation extensive de l'actualité. C'est bien entendu le cas des journalistes, qui sont environ 37 000 en France, et dont le rôle est de mettre en forme l'information pour la diffuser au public. On peut distinguer les journalistes responsables de la publication sur le web des journalistes publiant dans des journaux imprimés selon deux axes : l'absence de contraintes de temporalité et d'espace (Fenton, 2010). La possibilité de publier en ligne a en effet aboli l'une des principales limites des éditions papiers, à savoir la place limitée du médium physique pour exposer l'actualité, résumée de façon humoristique par Jerry Seinfield par la formule : « Il est étonnant de voir que le nombre d'événements qui se produisent chaque jour dans le monde remplissent exactement un journal ». Cette nouvelle liberté d'espace potentiellement infini a notamment permis une plus forte multiplicité des types de contenus, révolutionnant la pratique du journalisme (Gunter, 2003).

La contrainte d'espace ayant disparu, la principale restante est donc celle du temps alloué pour traiter chaque article. La concurrence entre les journaux *mainstream* étant sévère sur le web (Newman et al., 2016), la plupart des médias en ligne optent pour des publications très rapides, en rupture profonde avec le format quotidien pratiqué dans la presse écrite (Davies, 2009). Cagé et al. (2016) montrent que dans un cas sur quatre, une actualité diffusée pour la première fois par un média en ligne sera reprise dans les 4 minutes par un de ses concurrents. Cette recherche de rapidité peut aller jusqu'à la mise en place de directs lors desquels un éditeur publie, de façon quasi instantanée, de petits articles reprenant le déroulement d'un événement. Ces directs peuvent avoir lieu pour des événements sportifs, se substituant aux formats télévisés ou radios déjà exploités depuis de nombreuses années, mais également pour des événements politiques tels que des élections ou des débats télévisés. Dans ce dernier cas, les directs écrits peuvent remplir plusieurs rôles, tels que la retransmission brute des informations, mais aussi le commentaire, voire la correction. Cela a notamment été le cas lors des débats de l'élection présidentielle française de 2017, lors desquels différentes rédactions rapportaient, commentaient et vérifiaient (ou *fact-checkaient*) les propositions des candidats [7]. Ces directs peuvent alors être envisagés comme des sources d'information autonomes, ou comme des seconds écrans offrant des informations complémentaires au média principal, ici le débat télévisé. Ils tentent ainsi d'occuper une place jusqu'ici dominée par les réseaux sociaux, notamment Twitter, très largement utilisé comme second écran pour commenter tout type d'émissions. Ils sont également un moyen d'ajouter du contexte à l'actualité, d'expliquer ses implications et/ou ses références historiques, une pratique qui n'a cessé de se développer, au delà du simple *reporting*, depuis 1950 (Fink et Schudson, 2014). Néanmoins, cette accélération du rythme de publication est perçue comme négative par une partie de la profession. Ils sont 9,9 % à dénoncer cette pratique aux États-Unis selon une étude de Willnat et Weaver (2014).

5. www.alexa.com/siteinfo
6. chiffres relevés pour l'année 2017 (janvier-juillet)
7. www.ouest-france.fr/elections/presidentielle/direct-debat-presidentiel-regardez-le-duel-macron-le-pen-4965314

Un autre nouvel enjeu des rédactions consiste à répondre à la crise de confiance qui existe entre les citoyens et les organes de presse. En plus de l'accroissement des contacts des rédactions avec leurs lectorats, via l'embauche de *community managers* ou la mise en place de directs telle que mentionnée plus haut, se développe un nouveau type de journalisme : le *data journalism* ou journalisme des données. Celui-ci, s'il n'est pas uniquement voué à renouer un lien de confiance, permet davantage de transparence sur les sources utilisées (Coddington, 2015) et illustre mieux le processus intellectuel menant à la rédaction d'un article, notamment du fait qu'il s'accompagne plus aisément d'illustrations personnalisables et interactives (Bradshaw, 2014). Il s'accompagne également régulièrement de vérification d'informations (*fact-checking*), une pratique de plus en plus répandue aux États-Unis comme en Europe (Graves et Cherubini, 2016).

Confrontés à cette crise de confiance et aux difficultés économiques, les journalistes sont nombreux à rapporter des craintes quant à leur futur, avec 59,7 % d'entre-eux indiquant que le journalisme va « dans la mauvaise direction » (Willnat et Weaver, 2014). Il convient en effet de noter que, à l'exception du journal exclusivement en ligne Médiapart, les journaux peinent à adapter leur modèle économique à l'arrivée de l'information en ligne et à la baisse des revenus publicitaires (Newman et al., 2016). Cette situation complexe a mené, selon de récentes études, à une large augmentation du plagiat entre journaux (Cagé et al., 2017). Ce constat semble néanmoins à nuancer, une analyse partielle d'un corpus journalistique réalisée à des fins d'évaluation des systèmes que nous avons développés dans le cadre de cette thèse révélant que la quasi totalité des documents d'actualité rapportaient au moins une information originale (voir section 8.3.1). Face à ces nouvelles contraintes, les journalistes ont besoin d'outils leur permettant de vérifier efficacement la véracité des informations qu'ils reprennent (Silvia, 2001), mais aussi de trouver rapidement des éléments de contexte afin d'enrichir leurs publications et se différencier de leurs concurrents.

1.2 Outils disponibles et attentes des professionnels de l'information

Nous détaillons dans cette section les outils disponibles pour la consultation d'actualités, qu'ils soient dédiés au grand public ou aux professionnels. Nous continuons avec la description d'un protocole d'étude des attentes des professionnels de l'information sur leur outillage et en livrons les résultats.

1.2.1 Outils disponibles

De nombreux outils existent et sont disponibles pour le grand public afin de parcourir l'actualité. On peut les répartir en 5 principales catégories : les journaux en ligne, les agrégateurs de contenu, les blogs, les sites communautaires, et les lecteurs de flux RSS.

Les journaux en ligne sont en majorité issus de rédactions à l'origine consacrées au format papier et sont en France l'un des moyens privilégiés d'accéder à l'information (Newman et al., 2016). La plupart des journaux historiques mettent à disposition des articles en ligne accessibles gratuitement. Certains optent pour un modèle *freemium*, qui limite l'utilisateur selon différents axes si celui-ci n'est pas abonné. Parfois, il lui sera impossible de consulter des articles publiés plus de trois jours plus tôt. D'autres fois, il ne pourra consulter qu'un nombre d'articles limité dans le mois. Enfin, il peut n'avoir accès qu'à

des articles courts relatant les faits, les articles plus conséquents étant réservés aux abonnés. Rares sont les journaux ayant fait le pari du tout-payant, et la plupart utilisent un mélange des limitations explicitées ci-dessus afin d'encourager les utilisateurs à s'abonner, tout en tirant une partie de leurs bénéfices via les publicités visionnées par les utilisateurs non abonnés (Herbert et Thurman, 2007). Néanmoins, ces revenus devenant de plus en plus faibles au fur et à mesure des années, les rédactions, pourtant concurrentes, sont amenées à s'associer afin d'utiliser des plateformes publicitaires communes, telles que Skyline ou Gravity [8].

L'outil le plus largement utilisé dans le monde par le grand public est l'agrégateur d'actualités. Son rôle consiste à regrouper et à organiser l'information. La plupart des agrégateurs répartissent les articles au travers de grandes catégories (économie, international, local, ...) avant de les organiser en petits groupes discutant chacun d'un même événement. Deux modes d'accès à ces articles existent selon les systèmes. Dans le premier cas, un seul article par événement est proposé, tel que présenté en figure 1.1. Il s'agit généralement de l'article ayant été publié le premier. Dans le second cas, un ou plusieurs articles sont présentés pour un seul événement, et une liste des autres sources disponibles est rendue facile d'accès, comme montré en figure 1.2, sur la partie gauche. La seconde catégorie est légèrement plus sensible aux erreurs de regroupement d'articles. En effet, il arrive de trouver, dans la liste des articles censés discuter un même événement, des documents traitant en réalité de thématiques différentes. À l'opposé, la première catégorie d'agrégateurs ne laisse voir qu'une seule référence pour chaque événement. Les seules erreurs qu'on puisse y trouver consistent donc en une répétition de deux événements identiques, détectés comme différents à tort, et bénéficiant du même coup chacun d'un espace d'affichage tandis qu'ils auraient pu être regroupés. L'ensemble des productions journalistiques des sources considérées par ces agrégateurs n'est généralement pas disponible, et seules les actualités jugées les plus pertinentes, ou les plus intéressantes, sont proposées à l'utilisateur, souvent par le biais d'un profilage de ce dernier. La mesure de l'intérêt intrinsèque qu'un sujet d'actualité peut avoir a néanmoins été relativement peu étudié (Del Corso et al., 2005). Le reste des documents d'actualités, non proposés par les agrégateurs, demeure toutefois disponible via les moteurs de recherche des sites des organes de presse.

Il existe trois principaux acteurs sur ce marché des agrégateurs : Google News, MSN Actualités et Yahoo News. D'autres agrégateurs, spécialisés dans certaines langues ou certaines zones géographiques, sont également présents (Sapo, Mynet, Naver, ...). Il est intéressant de noter que Google News, l'un des principaux leaders présent dans 70 pays et en 35 langues (Calzada et Gil, 2016), a récemment mis à jour son agrégateur. En lieu et place de la liste brute d'articles discutant un événement, une catégorisation supplémentaire permet de distinguer les articles publiés depuis d'autres pays que ceux qui sont régulièrement cités et d'où l'information provient probablement. Cette nouvelle interface fournit également des liens vers des sites de vérification d'information (*fact checking*). Cette catégorisation supplémentaire vise à mieux informer le lecteur sur le type d'article qu'il s'apprête à lire si toutefois il venait à cliquer sur ce lien hypertexte. La comparaison des deux versions de l'agrégateur est illustrée en figure 1.2.

Les blogs sont très fréquemment hébergés par les sites des médias traditionnels. C'est notamment le cas du Monde [9], du Figaro [10], ou de Médiapart [11]. Ils sont alors dédiés aux

8. www.ouest-france.fr/economie/publicite-en-ligne-la-guerre-est-declaree-5132703
9. www.lemonde.fr/blogs/
10. www.lefigaro.fr/blogs/
11. www.blogs.mediapart.fr/

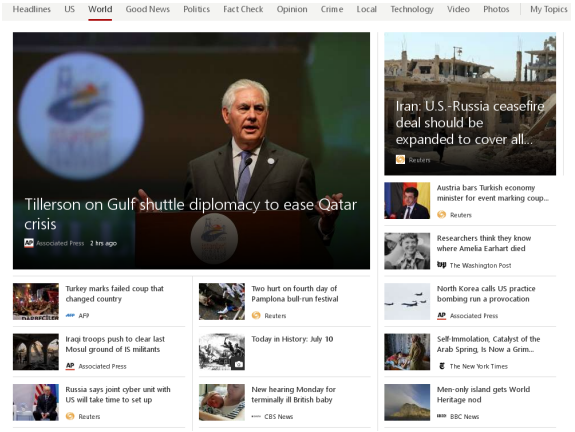


FIGURE 1.1 – Exemple d'un agrégateur d'actualités : MSN Actualités.

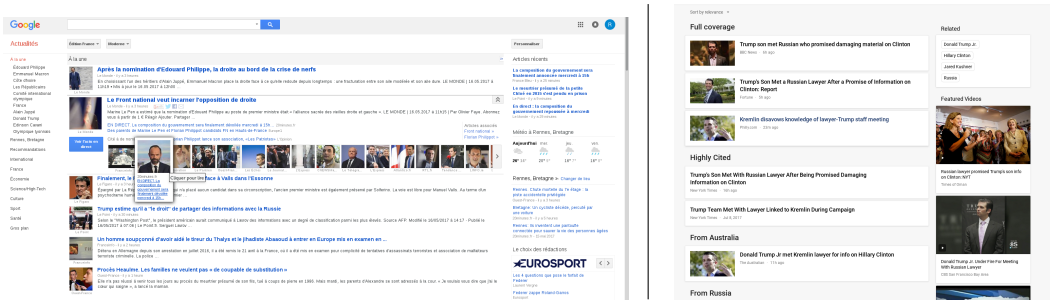


FIGURE 1.2 – Exemple d'un agrégateur d'actualités récemment mis à jour (droite) et sa version antérieure (gauche) : Google News.

rédacteurs du journal, qui peuvent s'y exprimer de façon plus personnelle, ou aux abonnés du journal, qui peuvent y trouver une tribune. Ils peuvent également être hébergés en dehors de toute structure éditoriale, et sont alors souvent initiés par des experts dans un but de vulgarisation ou de transmission du point de vue d'une profession ou d'une partie d'une profession (avocats [12], juges [13], policiers [14], ...). Leur rythme de publication est plus lent que les organes de presse classique, et les articles publiés y sont souvent plus longs, laissant une part plus importante à l'anecdote et à la subjectivité (Graves, 2007).

Les sites communautaires, tels Reddit, sont également largement utilisés par le grand public. Ces sites reposent sur une dynamique de production participative implicite (Doan et al., 2011; Weninger, 2014) (*implicit crowdsourcing*) afin de régler deux des problèmes les plus prégnants des agrégateurs d'actualités : le choix du contenu à mettre en avant et les sources d'information à utiliser. En effet, chaque individu présent sur ces sites peut soumettre un contenu qu'il juge intéressant sans avoir à choisir au sein d'une liste de sources prédéfinie. On pourra ainsi trouver aussi bien des articles de presse traditionnelle que des communications sur les réseaux sociaux, des articles de blog, ou des « unes » de journaux numérisées. En parallèle de ces soumissions, les participants au site sont invités à voter en faveur des contenus qu'ils jugent intéressants et en défaveur des contenus qu'ils jugent inintéressants. Un filtre est ainsi opéré directement par les utilisateurs, les contenus qui reçoivent beaucoup de votes favorables bénéficiant d'une large visibilité, et

12. www.maitre-eolas.fr
13. www.blog.francetvinfo.fr/judge-marie
14. www.police.etc.over-blog.net/

ceux qui reçoivent des avis négatifs étant masqués.

Enfin, les lecteurs de flux RSS, tels Awasu [15] ou NewsBlur [16], permettent à leurs utilisateurs de suivre une liste de médias de leur choix, et d'y mélanger blogs, presse traditionnelle et sites communautaires. S'ils demandent une phase d'amorce laborieuse, lors de laquelle l'utilisateur doit lister l'ensemble des sources qui l'intéressent, ils permettent ensuite de visualiser d'un coup d'œil les nouveaux articles publiés depuis une visite précédente. Ces lecteurs de flux disposent aussi parfois de fonctionnalités supplémentaires, telles que la génération automatique de résumés, la consultation hors-ligne, la consultation des sources citées dans les articles, ...

Les journalistes, en plus des outils disponibles pour le grand public, ont à leur disposition une ressource inestimable : l'accès à des archives journalistiques. Celles-ci peuvent être parcourues à l'aide de moteurs de recherche classiques, mais également par le biais de métadonnées annotées manuellement telles que les entités nommées mentionnées dans les documents (personnes, entreprises, lieux, ...), les dates de publication, les éditeurs... Les journalistes bénéficient également de l'aide d'archivistes, qui savent parfaitement décrire et manipuler les collections à leur disposition, et dont la mémoire permet de retrouver aisément des articles liés à un événement d'intérêt. Lorsque l'information qu'ils recherchent n'est pas disponible dans leurs archives, ou bien qu'ils cherchent à inclure davantage d'éléments de contexte, les journalistes se tournent vers des ressources externes telles les encyclopédies (22,2 % des journalistes utilisent Wikipedia dans leur travail selon Willnat et Weaver (2014)). Ils utilisent également assez largement les réseaux sociaux, et notamment Twitter, qui est utilisé par 53,4 % des journalistes selon la même étude. Ceci s'explique en partie par le fait que de plus en plus de médias utilisent des tweets pour illustrer leurs articles.

1.2.2 Protocole d'étude des besoins des professionnels de l'information

Comme nous l'avons décrit précédemment, de nombreux outils existent pour l'exploitation de données multimédias à caractère journalistique. Pourtant, les professionnels que nous avons rencontrés n'utilisent pas de systèmes sophistiqués issus de la recherche mais plutôt des outils simples et grand public, tels que les moteurs de recherche. Il est alors naturel de se demander pourquoi les rédactions ne sont pas davantage équipées, d'autant plus que la profession utilise déjà largement de nouvelles technologies comme les réseaux sociaux. Nous avons donc mené, en collaboration avec le CRPCC et dans le cadre du projet LIMAH, une étude fondée sur de longs entretiens avec des professionnels des médias et visant à comprendre quelles fonctionnalités sont attendues d'un outil leur permettant une plus grande efficacité dans leur travail de rédaction. Nous avons voulu ces entretiens orientés utilisateurs plutôt que technologie. Ainsi, nous, chercheurs, nous sommes confrontés aux besoins des professionnels, plutôt que de confronter les professionnels aux outils que *nous* leur estimons utiles. Ce travail, bien que situé dans un chapitre initial de ce manuscrit, constitue une première contribution de notre thèse. Décrivant un travail amont nécessaire, il a été placé ici à dessein. Les résultats, résumés plus bas, ont fait l'objet d'une publication à l'international (Gravier et al., 2016). Ces entretiens sont fondés sur la norme ISO 9241-210 (*Human-centred design for interactive systems*) qui identifie quatre principales étapes à la conception d'applications :

1. comprendre et caractériser le contexte d'utilisation ;

15. www.awasu.com/
16. www.newsblur.com

2. identifier les besoins utilisateurs ;

3. prototyper des solutions ;

4. évaluer ces solutions.

Nous nous concentrons sur les trois premières étapes, et nous focalisons sur deux cas d'utilisation : la recherche d'information, qui consiste à répondre à une question précise, et l'exploration d'actualités, qui consiste à appréhender un sujet d'intérêt et à en comprendre les éventuelles ramifications. Dans les deux cas, le contexte des données disponibles est multimédia, multimodal et multisource. Il est composé d'articles de presse, d'émissions télévisuelles et radiophoniques en grand nombre, ainsi que de contenus utilisateurs associés tels que les commentaires, tweets et réactions.

Nous avons conduit en collaboration avec le CRPCC des entretiens sur un mois avec 13 professionnels des médias représentant 3 professions, à savoir des journalistes, des agents de presse et des *community-managers*. Tous analysent la presse de façon quotidienne avec des obtectifs différents. Ces professions ont été choisies afin de prendre en compte la diversité des pratiques liées à la consultation d'actualités. Le nombre de sujets a été jugé suffisant, 5 à 10 personnes suffisant généralement à soulever la majorité des problématiques d'utilisabilité (Nielsen et Landauer, 1993). Chaque entretien a été divisé en trois parties. Un questionnaire a d'abord été utilisé afin de mieux connaître le profil des personnes interrogées, en particulier leur appétence pour les nouvelles technologies et leur relation avec la presse. L'entretien en lui-même, dirigé par un ergonome, a commencé par une analyse des activités et pratiques actuelles de la personne interrogée. Enfin, les sujets ont été questionnés sur l'acceptabilité des fonctionnalités pouvant répondre aux problématiques posées dans le cadre de leur emploi. Chaque entretien a duré entre 1 heure et 1 heure 30.

Les 13 personnes interrogées sont majoritairement des hommes (9/13) et se répartissent en 4 agents de presse, 7 journalistes, et 2 *community-managers*. Leur âge moyen est de 32 ans, avec une expérience professionnelle moyenne de 5 ans. Nous avons mesuré plusieurs facettes de leur profil technologique. Leur empressement à tester des technologies innovantes a obtenu un score moyen de 5 sur une échelle de Likert allant de 0 (fort désaccord) à 10 (fort accord), à l'exception des *community-managers* qui ont obtenu un score moyen de 8,5. Leurs compétences d'utilisation d'Internet et de recherche d'information vont du moyen pour les agents de presse (6 sur l'échelle de Likert) à haut pour les *community-managers* (10 sur l'échelle de Likert), les journalistes se situant au milieu des deux autre groupes. Tous les professionnels utilisent largement les réseaux sociaux, confirmant l'étude de Willnat et Weaver (2014). Les réseaux les plus utilisés sont Facebook (100 % des interrogés) et Twitter (91,7 %). 58 % des interrogés passent plus de 5 heures par jour sur les réseaux sociaux. Les moteurs de recherche sont fréquemment utilisés par tous les groupes étudiés.

Nous avons également mesuré quelles sources d'information sont utilisées par ces professionnels. Les actualités radiophoniques sont les plus régulièrement utilisées par les agents de presse et les journalistes, suivies par les journaux écrits. Les sites d'agrégation d'actualité sont relativement peu utilisés, à l'exception des journalistes qui sont 3 sur 7 à les visiter au moins une fois par jour.

1.2.3 Acceptabilité des fonctionnalités pour les professionnels de l'information

La suite des entretiens a eu pour objectif d'analyser l'acceptabilité, c'est-à-dire l'utilité perçue d'un outil ou d'une technologie. Un ensemble de fonctionnalités, fréquemment développées dans le cadre de projets de recherche, ont ainsi été répertoriées et présentées aux professionnels interrogés. Pour chaque fonctionnalité proposée, une échelle de Likert allant de 0 à 10 a permis d'évaluer l'utilité perçue. Les verbatims des entretiens se sont également révélés précieux afin de déceler ce qui était attendu des outils proposés. Il est important de noter que ces tests d'acceptabilité ont été réalisés sans fournir de système aux expérimentateurs. Les fonctionnalités ont été seulement décrites de façon orale. La personne testée devait donc imaginer comment ces fonctionnalités pourraient être utilisées dans son travail, sans être confrontée à un système précis dans lequel les écueils classiques d'utilisabilité sont présents (clarté de l'interface, accessibilité des fonctionnalités, ...). Plusieurs *mock-ups* ont également été présentés afin d'illustrer les fonctionnalités proposées (voir figure 1.4). Les fonctionnalités étudiées sont regroupées au sein des 4 catégories suivantes :

1. contenu et mots-clés ;
2. réseaux sociaux et opinions ;
3. liens et recommandations ;
4. abstractions du contenu et accès rapides.

La figure 1.3 présente l'utilité perçue pour chacune de ces grandes catégories. Nous développons ces résultats ci-après. La première catégorie regroupe l'affichage du contenu, y compris sous forme de transcriptions dans le cas de vidéos ou d'émissions radiophoniques. Elle propose également de lister les mots-clés et entités nommées présentes dans le document, ainsi que l'affichage d'un nuage de mots. 6 personnes sur 13 ont jugé que l'affichage des transcriptions était utile, et 3 ont mentionné qu'une telle fonctionnalité leur ferait gagner du temps dans leur travail. L'affichage de nuages de mots a quant à lui été perçu négativement, avec 3 personnes déclarant qu'ils ne représentaient pas d'intérêt et 2 autres qu'ils n'étaient pas pratiques. À l'opposé, les mots-clés sont perçus positivement, et particulièrement leur ordonnancement par fréquence ou importance dans le texte considéré. Mettre en avant les noms propres a été perçu comme utile ou intéressant par 5 personnes, mais seule une d'entre elles a déclaré son intérêt pour un lien direct de ces noms propres vers les biographies correspondant. Mettre en avant les lieux discutés dans les documents a également, de façon surprenante, été jugée comme inutile par 3 personnes, 3 autres indiquant qu'une carte affichant ces lieux serait davantage appropriée. Tous ces retours sur ces différentes fonctionnalités ont été largement conditionnés selon deux critères supplémentaires : leur nécessaire précision, et la fiabilité des sources d'information utilisées. Bien que plusieurs fonctionnalités puissent être développées de façon fiable étant donné l'état de l'art, d'autres, telles que l'affichage de la transcription automatique, risquent d'être jugées comme inutiles si trop inexactes.

La deuxième catégorie étudiée concerne les réseaux sociaux et l'opinion. Elle comprend des fonctionnalités de détection de la valence de l'opinion exprimée (positive, négative ou neutre) à un moment donné ou sur la durée, mais également des outils plus précis tels que la détection des sentiments (Fraisse et Paroubek, 2014) (colère, surprise, peur, ...), qui ne sont à l'heure actuelle que partiellement étudiés. Des fonctionnalités à grains fins sont également envisagées comme la description des différents documents

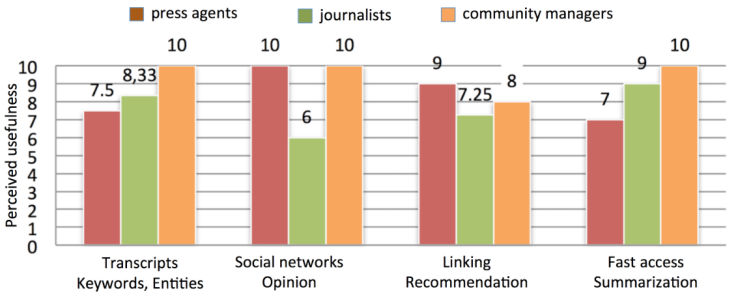


FIGURE 1.3 – Utilité perçue des fonctionnalités en fonction de la profession.

FIGURE 1.4 – Mock-up présenté aux expérimentateurs pour la catégorie "liens et recommandations".

en aspects et la détection des marqueurs d'opinions associés à chacun de ces aspects. Ces marqueurs peuvent ensuite être regroupés afin d'obtenir une vue synthétique des opinions relayées sur les réseaux sociaux. Cette catégorie de fonctionnalités a été perçue comme intéressante et utile, 5 interrogés mentionnant son utilité et 3 mentionnant un possible gain de temps dans leur travail. C'est notamment le cas pour les agents de presse, au contraire des journalistes qui sont plus sceptiques sur l'utilité et l'efficacité de ce genre d'information. L'affichage de l'évolution de l'opinion au cours du temps a été perçu comme utile par 5 personnes. La consultation de l'opinion semble plus adaptée aux professionnels pour un événement précis tandis que les sentiments seraient plus adaptés à une vision globale. De façon logique, ces analyses d'opinions sont perçues comme peu pertinentes lorsqu'appliquées aux médias eux-mêmes plutôt qu'aux réactions des utilisateurs. Il est intéressant de noter que 5 personnes ont suggéré de rendre les différentes analyses d'opinions filtrables en fonction de leur source (par réseau social, par mot-clé, par importance, …). Globalement, il semble exister chez les professionnels une grande demande pour des outils d'analyse fine des sentiments sur les réseaux sociaux, qui vont au-delà de la mesure de la valence. La disponibilité d'une explication des résultats des algorithmes de détection d'opinion est également réclamée par les professionnels, au travers d'exemples représentatifs et de mesures de certitude.

La troisième catégorie considérée concerne la création de liens explicites entre les documents d'une collection. La figure 1.4 présente le *mock-up* proposé aux interrogés pour

cette catégorie. Il peut s'agir de liens de type recommandation, d'ordonnancement temporel ou de regroupement d'articles discutant un même événement. La décision de lier deux documents peut répondre à de nombreux critères tels qu'une thématique commune, les mêmes personnages impliqués, un contenu similaire... Nous avons demandé aux personnes interrogées de classer ces types de liens selon leur utilité perçue. Nous avons obtenu l'ordre suivant : une thématique commune, des mots-clés partagés, une date identique, une localisation proche. L'utilité de l'ordonnancement temporel a néanmoins été régulièrement mentionnée lors des entretiens. L'explication des liens a été perçue comme utile, une explication fine étant plébiscitée (score de 8,5) par rapport à un typage grossier des liens (score de 6,3). Limiter la redondance en regroupant les documents similaires n'a pas été jugé comme étant une fonctionnalité nécessaire. Seuls les agents de presse ont plébiscité les fonctionnalités de « regroupement d'articles similaires » et de « mise en avant des documents représentatifs » (scores de 10). Deux personnes ont suggéré de mettre en avant des documents non issus de la presse, mais ayant servi de source lors de la rédaction d'articles.

La quatrième et dernière catégorie correspond à l'accès rapide à l'information et à la visualisation d'une collection d'actualités dans son ensemble ou en partie. Les fonctionnalités classiques de moteur de recherche et de table des matières ont été proposées, ainsi que plusieurs méthodes d'accès rapide telles que la présentation d'une ligne de temps (*timeline*), un nuage de mots cliquables, le résumé automatique des documents similaires. Les retours obtenus sur cette catégorie montre un global manque d'intérêt en dehors de fonctionnalités classiques de moteur de recherche et de *timeline*. En particulier, la génération de résumés automatiques a été jugée non pertinente, non pas pour son intérêt intrinsèque, mais par crainte d'une performance insuffisante, les professionnels estimant qu'il était difficile de faire confiance aux résumés générés par une machine, alors même que des études ont prouvé l'intérêt de ce genre d'outils pour la rédaction d'articles (McKeown et al., 2005). Ces remarques mettent en avant le fait que la tâche consistant à résumer un ou plusieurs documents est le plus souvent une tâche subjective, dans laquelle l'auteur du résumé choisira de traiter un angle plutôt qu'un autre. Bien que la génération automatique de résumés orientés utilisateurs fasse l'objet de travaux de recherche depuis quelques années (Lin et Hovy, 2000; Daumé III et Marcu, 2006; Hennig et Labor, 2009), les résumés générés obtiennent la plupart du temps des résultats inférieurs aux résumés dits « neutres », et il est souvent complexe pour l'utilisateur d'indiquer l'angle qu'il souhaite donner à ces derniers.

Conclusion

Comme nous avons pu le décrire dans ce chapitre, la consultation d'actualités soulève de nombreux enjeux et des besoins différents pour le grand public et pour les journalistes. Si le grand public paraît être globalement satisfait des outils à sa disposition, il doute de plus en plus largement de la fiabilité des médias à sa disposition. Les journalistes semblent majoritairement avoir choisi de renforcer leur crédibilité via un contact plus important et plus direct avec leur lectorat. Ils organisent des suivis en direct de certains événements, lors desquels leurs lecteurs peuvent interagir et poser leur questions, embauchent des *community managers* afin d'occuper l'espace des réseaux sociaux et de se montrer plus proches des préoccupations de ceux qui les lisent. Les fournisseurs d'agrégats de contenus, quant à eux, semblent davantage s'orienter vers une présentation de plus en plus fine de l'actualité. Les plus récentes mises à jour de leurs systèmes offrent

ainsi des liens vers des sites de vérification des faits (*fact checking*), et distinguent les articles de presse des éditoriaux.

Un aspect reste néanmoins négligé : le besoin de mettre en relation différentes informations, afin de comprendre comment elles s'imbriquent les unes avec les autres dans un tout cohérent. C'est ce dernier aspect qui est principalement abordé au sein de cette thèse. L'étude que nous venons de décrire permet ainsi de confirmer l'intérêt d'une structure proposant des liens entre différents documents d'actualités. Cet intérêt sera confirmé en aval, par le biais de nouvelles études utilisateurs mettant à contribution des professionnels de l'information (voir chapitre 8).

Chapitre 2

Outils scientifiques pour la consultation et la structuration de collections d'actualités

Introduction

Plusieurs domaines de recherche se sont penchés sur la consultation d'actualités. C'est le cas de la recherche en traitement des langues, qui s'est intéressée à la structuration de collections composées d'articles de presse, notamment en exploitant la notion d'événement (Van Hage et al., 2011), unité élémentaire de l'information. Ces événements sont la plupart du temps ordonnés de façon temporelle ou causale, regroupés en sujets ou thématiques, et font l'objet de nombreux traitements annexes comme l'extraction d'entités nommées ou la génération automatique de résumés. C'est également le cas de la communauté multimédia, au travers de la recherche d'information multimédia et plus particulièrement, depuis quelques années, de l'hyperliage. Cette communauté s'intéresse également de près à la construction de systèmes aux nombreuses fonctionnalités permettant de parcourir des collections vidéos, souvent liées à l'actualité.

Dans ce chapitre, nous présentons un état de l'art des méthodes facilitant la consultation d'actualités dans les domaines du traitement des langues et du multimédia. Ces méthodes peuvent se diviser en deux groupes : celles qui tentent de masquer la multiplicité des informations disponibles par des approches de regroupement (*clustering*) d'articles similaires, et celles qui proposent une structuration des collections d'actualités permettant de naviguer d'article en article en suivant des liens temporels ou thématiques. Après avoir décrit ces deux catégories, nous détaillons quelques systèmes complets, constitués de plusieurs composants (transcription, indexation, résumé automatique, ...) et développés au sein de projets de recherche.

2.1 Groupement d'articles similaires

Une première approche permettant l'appréhension de grandes collections d'actualités consiste à regrouper les articles similaires. Selon la granularité des groupes créés, on

peut obtenir une collection organisée en grandes catégories ou en petits groupes très homogènes discutant le même événement. Nous discutons dans cette section de ce large spectre.

2.1.1 Catégorisation

Étant donné la grande quantité et la variété des informations disponibles chaque jour, il semble intuitif de chercher à rassembler les articles similaires au sein de groupes plus ou moins grands. Une première étape de ce regroupement consiste à catégoriser les articles en thématiques. Ces catégories peuvent être grossières, et correspondre aux sections que l'on trouve dans un journal papier ou un agrégateur (national, international, local, sports, ...), ou fines, telles que le standard de l'International Press Telecommunications Council [1] (IPTC), qui consiste en 17 grandes catégories (politique, religion, économie, ...) et en plusieurs niveaux de sous-catégorisation (élections nationales, fêtes religieuses, chômage, ...).

Dans le premier cas, c'est-à-dire si le nombre de catégories se réduit à un petit ensemble, il est possible d'obtenir de très bons résultats via des approches entièrement automatiques. Ainsi, Bracewell et al. (2009) parviennent à atteindre environ 95 % de précision et de rappel (micro et macro), sur une dizaine de catégories, en anglais et en japonais, grâce à un système d'apprentissage fondé sur la fréquence des groupes nominaux et un calcul des probabilités d'appartenance à chacune des catégories. Des résultats similaires peuvent être obtenus avec d'autres algorithmes d'apprentissage automatique, l'un des plus utilisés pour cette tâche étant les machines à vecteurs de support (*support vector machines*, ou SVM) (Joachims, 1998; Krishnalal et al., 2010). Des approches purement statistiques peuvent également être employées telles que le *Naïve Bayes* (Diriye et al., 2010).

Le second cas, plus complexe, consiste à attribuer automatiquement plusieurs étiquettes, par exemple celles proposées par la taxonomie de l'IPTC. L'Agence France Presse (AFP) ainsi que de nombreux autres organes de presse utilisent cette taxonomie, chacun de leurs articles étant manuellement catégorisé, et plusieurs catégories et sous-catégories pouvant être attribuées à un même article (Cagé et al., 2016). Cette catégorisation peut être apprise automatiquement via un apprentissage supervisé, mais les multiples étiquettes possibles rendent sa mise en place et son évaluation complexes. Bacan et al. (2005) se sont attaqués à ce problème, et obtiennent des résultats relativement corrects grâce à une approche des K plus proches voisins (*K nearest neighbors* ou *K*-NN) (87 % de précision), mais ne considèrent que les 17 étiquettes de plus haut niveau, et simplifient le problème davantage en attribuant une unique étiquette, évaluée positivement lorsqu'elle fait partie des multiples catégories assignées à l'article. Cette faible fiabilité ne permet pas la mise en place de systèmes entièrement automatiques, qui ne font aujourd'hui qu'aider l'annotation manuelle en attribuant une première étiquette. Néanmoins, l'approche qu'ils ont choisie leur permet en fait de proposer une liste d'étiquettes potentielles et pas uniquement la plus probable, ce qui peut constituer une aide déterminante pour l'étiquetage manuel sans toutefois pouvoir le remplacer totalement.

2.1.2 Regroupement statique

Si la catégorisation permet de regrouper les articles discutant d'une même large thématique, elle se révèle insuffisante lorsque l'on considère plusieurs sources d'informa-

1. show.newscodes.org/index.html?newscodes=medtop&lang=en-GB&startTo=Show

tions (différents médias, blogs, ...) sur des durées longues. En effet, dans ce cas, les catégories peuvent contenir plusieurs milliers de documents, rendant leur exploration complexe et laborieuse. C'est pourquoi, comme vu précédemment, le regroupement d'articles en petits groupes, ou *clustering*, est très largement utilisé par les agrégateurs d'actualités. Il consiste généralement à regrouper les articles discutant d'un même événement (par opposition aux catégories, plus larges), et publiés dans un laps de temps restreint, de l'ordre de la journée. On parle de regroupement statique, par opposition au regroupement dynamique, dans le cas de collections fixées, c'est-à-dire celles dans lesquelles aucun nouveau document n'est inséré.

Afin de pouvoir regrouper les articles similaires, une définition de la similarité s'impose. Bien que ce concept soit intuitif à l'humain, il est complexe à définir précisément. La similarité entre deux documents peut en effet revêtir de nombreux aspects. Elle peut être thématique (ils abordent des sujets proches), sémantique (ils signifient la même chose), lexicale (ils emploient le même vocabulaire)... D'un point de vue pratique, le processus de calcul de similarité entre deux documents est invariablement constitué de deux étapes : dans un premier temps, on cherche à obtenir une représentation lexicale, sémantique, thématique ou autre, des documents, et on use, dans un second temps, de mesures de similarité fondées sur ces représentations afin de déterminer un score de similarité pour chaque paire de documents. Ces deux étapes peuvent être réalisées de multiples façons, comme décrit par Gomaa et Fahmy (2013). La représentation et la mesure de similarité les plus utilisées sont lexicales. Il s'agit du tf-idf (Salton et Buckley, 1988) et ses dérivés et la mesure cosinus (Strehl et al., 2000). La représentation tf-idf correspond à un espace vectoriel de grandes dimensions. Les documents sont représentés sous forme de vecteur par les mots qu'ils contiennent, chaque mot du vocabulaire correspondant à une dimension. D'autres représentations vectorielles, généralement fondées sur un principe de similarité contextuelle (Mikolov et al., 2013) – corrélée avec la similarité sémantique (Miller et Charles, 1991) –, ont été récemment introduites (Kiros et al., 2015).

Une fois la représentation et la mesure de similarité entre paires de documents établies, on peut s'attacher à trouver les groupes pertinents au sein de la collection. Le problème consistant à trouver quels groupes d'articles sont pertinents au sein de cet espace peut être formulé sous forme d'optimisation : il s'agit de maximiser la similarité entre les éléments appartenant à un même groupe et de minimiser la similarité entre éléments de groupes différents. C'est sur ces bases que repose *k-means*, un algorithme efficace même à grande échelle (Sculley, 2010), et qui groupe de façon itérative les documents proches, jusqu'à atteindre un état stable. Néanmoins, cet algorithme est connu pour être très dépendant de ses paramètres d'initialisation (Bradley et Fayyad, 1998), et atteint souvent des optima locaux plutôt que globaux. Il peut être adapté à des systèmes dits « en ligne », c'est-à-dire mis à jour de façon continue avec l'arrivée de chaque nouvel article (Azzopardi et Staff, 2012). D'autres techniques d'optimisation, plus robustes, existent telles que le regroupement à base de recherche d'harmonie (*Harmony Search CLUSTtering* ou HSCLUS) (Forsati et al., 2013).

Le regroupement hiérarchique est une autre méthode permettant de construire les groupes de façon itérative, en commençant par grouper les deux documents ayant le score de similarité le plus élevé. Le processus est répété, à la différence que les éléments à lier ne sont plus nécessairement des documents, mais peuvent être des groupes de documents issus d'une itération précédente. Trois grandes approches sont alors possibles pour obtenir un score de similarité entre deux groupes de documents : considérer le score des deux documents les plus similaires entre deux groupes (il y a alors risque de dérive et for-

mation de grands groupes), le score des deux documents les moins similaires (on obtient généralement davantage de petits groupes), ou la moyenne des scores des documents de chacun des groupes (dans l'esprit plus proche de l'algorithme *k-means*) (Hatzivassiloglou et al., 2000). Le regroupement hiérarchique est également très largement utilisé pour le regroupement d'actualités (McKeown et al., 2002; Nallapati et al., 2004), y compris dans des sous-domaines comme l'actualité financière (Dai et al., 2010). Il repose néanmoins sur un seuil difficile à déterminer automatiquement (nombre de groupes ou seuil de similarité minimale) (Salvador et Chan, 2004). Un avantage certain de cette approche est qu'elle permet l'affichage direct des hiérarchies obtenues, et donc de plusieurs niveaux de groupements (sous-groupes). Ces hiérarchies sont néanmoins difficiles à visualiser dans de grandes collections (Rennison, 1994).

2.1.3 Regroupement dynamique

Le regroupement dynamique est semblable au regroupement statique, à ceci près qu'il se concentre sur des collections qui évoluent et auxquelles de nouveaux articles sont ajoutés en continu. Si ces nouveaux documents arrivent par lots, par exemple via une mise à jour quotidienne, les approches décrites en section 2.1.2 peuvent suffire. Néanmoins, lorsque la mise à jour est continue, c'est-à-dire que l'on souhaite traiter un document dès qu'il est publié, deux nouvelles tâches apparaissent : la détection automatique d'un nouveau sujet encore jamais rencontré (*first topic detection* ou FTD), et le choix du groupe auquel rattacher un nouvel article (*topic tracking*). Ces deux problématiques, ainsi que quelques autres, ont été étudiées dans le cadre des campagnes d'évaluation Topic Discovery and Tracking (TDT) (Allan et al., 1998).

La détection d'un nouveau sujet est le plus souvent réalisée à l'aide de seuils. Voici par exemple l'approche utilisée par la *Carnegie Mellon University* pour cette tâche (Carbonell et al., 1999). Lorsqu'un nouveau document arrive, il est comparé à l'ensemble des groupes déjà existants suffisamment récents via des méthodes classiques de comparaison vectorielle de documents, à savoir une représentation tf-idf et une mesure cosinus. Seul un facteur de diminution du score en fonction de l'ancienneté est ajouté à cette comparaison. Si la similarité du nouveau document avec l'un des groupes est suffisamment élevée (au-dessus d'un seuil fixé manuellement), le nouveau document est considéré comme appartenant à ce groupe, y est ajouté, et n'enclenche donc pas la création d'un nouveau groupe. Si la similarité est trop faible (en-dessous du seuil), le nouveau document est considéré comme un nouvel événement. Dans ce cas, il devient l'unique composant d'un nouveau groupe, qui est ajouté à la liste des groupes déjà existants. Si le nombre de groupes déjà existants est trop élevé, le groupe le plus ancien (celui qui n'a pas reçu de nouvel article depuis le plus longtemps) est supprimé. Des difficultés apparaissent néanmoins lorsque les documents traités couvrent plusieurs semaines ou plusieurs mois (Allan et al., 2000a). En effet, le vocabulaire utilisé pour décrire un même événement peut évoluer, rendant les mesures de similarité par représentation de surface peu fiables (Tannier et al., 2012).

Le choix du groupe auquel rattacher un nouvel article reprend les mêmes problématiques de similarité évoquées précédemment. Il s'agit finalement de trouver le groupe le plus similaire au nouvel article. Pour ce problème, l'utilisation d'un algorithme de plus proches voisins (*k nearest neighbors* ou *K*-NN) a été l'approche la plus performante (Allan et al., 1998). Des arbres de décision, qui ont permis l'extraction de caractéristiques plus riches que les représentations *K*-NN (*e.g.*, présence de certains mots en début de paragraphe, collocation de bigrammes, ...), ont été utilisés lors d'une édition de TDT suivante, avec des résultats similaires ou supérieurs à *K*-NN (Carbonell et al., 1999).

Le regroupement, qu'il soit statique ou dynamique, est particulièrement adapté à la consultation rapide des informations publiées au cours des dernières heures, mais devient moins pratique au fur et à mesure que la période considérée s'étend. En effet, si quelques dizaines de groupes permettent de disposer d'une vue d'ensemble des articles publiés par les médias professionnels en ligne au cours des dernières heures, il faudrait plusieurs milliers de groupes pour présenter l'ensemble des actualités d'un mois, plus encore si l'on étendait la période considérée à plusieurs années. Bien que quelques systèmes fondés sur ces techniques aient vu le jour, ils semblent difficiles à appréhender pour le grand public (Frey et al., 2001). Pour ce genre de corpus, il est plus pertinent de chercher à structurer la collection plus finement.

2.2 Structuration de collections

Les groupes de documents obtenus par les approches décrites précédemment peuvent être raffinés afin d'expliciter des liens temporels ou causaux entre documents. Dans cette section, nous explorons différentes méthodes permettant un tel raffinage et en donnons les points forts et les points faibles.

2.2.1 Structuration chronologique

La structuration chronologique correspond à organiser une collection selon un axe temporel. Elle est le plus souvent un traitement supplémentaire réalisé après les étapes de classification et de regroupement décrits précédemment (Ahmed et al., 2011). Elle est particulièrement adaptée à deux types d'actualités : les *breaking news*, qui mènent à la publication de nombreux articles révélant les derniers rebondissements d'un événement, et les événements se déroulant sur plusieurs jours, semaines, ou mois (élections, jeux olympiques, ...). Deux approches peuvent alors cohabiter : résumer l'ensemble des articles en construisant une *timeline* qui reprend les événements majeurs (Yan et al., 2011), ou organiser l'ensemble des documents sur une ligne temporelle (Allan et al., 2000b; Swan et Allan, 2000). Dans le second cas, les métadonnées associées aux articles sont généralement suffisantes pour réaliser un ordonnancement temporel (Mori et al., 2006), les dates de publication correspondant souvent aux dates des événements discutés. Néanmoins, certains articles mentionnent différents événements qui sont survenus à des périodes différentes. Dans ces cas, la détection des dates, leur normalisation et leur réorganisation chronologique apparaîssent indispensable (Muller et Tannier, 2004).

La campagne d'évaluation TempEval, qui a connu trois éditions (UzZaman et al., 2012; Verhagen et al., 2007, 2010), s'est attaquée à ce problème en cherchant notamment à repérer les mentions temporelles, mais également les liens temporels entre événements (*e.g.*, précédence, simultanéité, ...). Une fois toutes ces relations établies, il devient possible de reconstruire une ligne temporelle récapitulant tous les événements. Ces structures chronologiques peuvent ensuite être utilisées, soit à des fins de présentation à l'utilisateur (voir la figure 2.1), soit dans le but de générer un unique document résumant les événements principaux qui la composent (Tannier et Vernier, 2016; Binh Tran et al., 2013). Néanmoins, elles ne rendent pas compte de la complexité de certains événements, qui nécessitent une structuration non linéaire afin d'être correctement appréhendés.

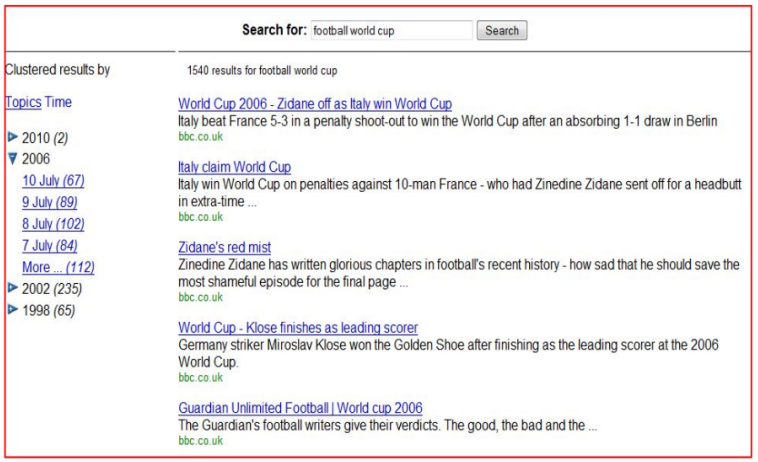


FIGURE 2.1 – Exemple d'organisation chronologique d'une collection (Alonso et al., 2009).

2.2.2 Fils d'actualités

L'étude de suites d'événements reliés les uns aux autres a progressivement amené à considérer une organisation plus riche que le regroupement ou l'ordonnancement temporel (Makkonen, 2003). L'organisation en fils d'actualités propose de regrouper les documents d'actualités au sein d'une structure arborescente rendant mieux compte des relations entre événements. L'apparition d'un nouvel événement générera ainsi un nouvel arbre, dont la racine sera le premier document mentionnant cet événement. Les documents suivants, qui décrivent les implications de cet événement, ne sont alors plus ordonnés seulement selon une ligne de temps, mais selon des branches temporelles. En effet, d'un même événement peuvent découler différentes conséquences, chacune d'entre elles faisant l'objet d'un suivi distinct dans l'actualité. Étant donné que plusieurs articles peuvent discuter un même événement sans forcément nécessiter la création d'un nouvel embranchement, cette méthode peut s'appliquer à des groupes d'articles. Nallapati et al. (2004) ont introduit cette problématique et annoté une partie des corpus de TDT-2 et TDT-3 afin d'obtenir les dépendances entre chacun des sujets abordés. La figure 2.2 donne un exemple d'arbre issu du corpus de TDT-3 et annoté par Nallapati et al. (2004). Dans leur article, les auteurs proposent des méthodes à base de similarité de surface et d'ordonnancement temporel afin de lier un nouveau sujet à son parent le plus similaire, c'est-à-dire au sujet publié plus tôt auquel il ressemble le plus.

Cette structuration en arbre a également été appliquée à des collections vidéo, aboutissant à des visualisations arborescentes qui facilitent l'exploration d'une thématique (Ide et al., 2004, 2012). D'un point de vue utilisateur, cette approche permet une meilleure compréhension de l'enchaînement des événements, qui est parfois difficile à appréhender lorsque les sujets sont considérés de manière indépendante. Elle autorise également la création de liens entre différents événements qui découlent les uns des autres, et donc de réunir les documents discutant ces événements au sein d'entités plus larges, permettant d'explorer plus facilement une collection de grande taille (de Rooij et Worring, 2010). Il est néanmoins impossible avec cette représentation de proposer une lecture multifactorielle d'un événement, qui peut être causé par plusieurs événements antérieurs.

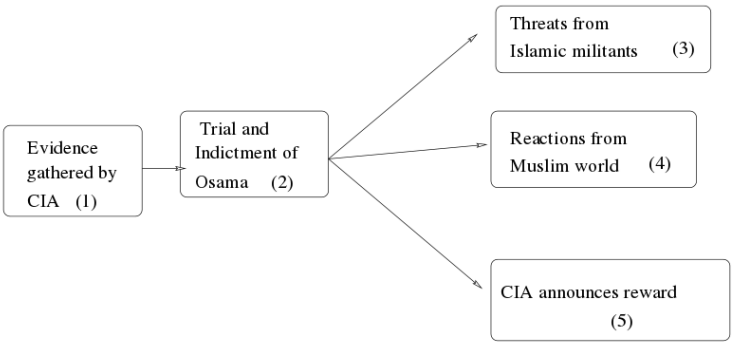


FIGURE 2.2 – Exemple de fil d'actualités issu de TDT-3 (Nallapati et al., 2004).

2.2.3 Graphes d'actualités

La structuration d'actualités en graphes pousse l'idée des fils d'actualités en permettant par exemple d'exprimer le fait qu'un événement puisse être causé par une combinaison d'événements antérieurs. Là où la structure en arbre ne permet à chaque événement de n'avoir qu'un unique parent, la structure en graphe abolit cette contrainte. Les graphes sont très largement utilisés dans de nombreuses applications du traitement automatique des langues (Nastase et al., 2015). Un graphe G est défini comme un ensemble de nœuds V et un ensemble d'arcs E tel que : $\forall e \in E, e = (v_i, v_j)\, v_i, v_j \in V^2$. Ici, les nœuds peuvent correspondre à des documents, articles ou vidéos d'actualités, ou bien à des sujets, groupes de documents discutant un même événement. Les arcs peuvent, comme pour le suivi d'actualités décrit par (Nallapati et al., 2004), correspondre à une relation de cause à effet, ou bien être considéré comme des liens thématiques, de sources, de recommandations, de lieux communs, ... La granularité de ces méthodes peut également être plus fine que le document. Ainsi, Glavaš et Šnajder (2014) extraient des documents les prédicats correspondant à des événements élémentaires, puis organisent ces prédicats selon leurs liens de causalité.

Choudhary et al. (2008) ont été parmi les premiers à proposer cette approche, en construisant un graphe dans lequel les documents mentionnant les mêmes personnes sont liés. Les événements rassemblant différentes personnes (*e.g.*, une rencontre entre deux présidents) sont connectés aux événements discutant individuellement de chacune de ces personnes. Ils proposent un graphe dirigé, dans lequel chaque arrête va d'un événement ancien à un événement plus récent. Le graphe est donc conçu pour être parcouru selon un paradigme chronologique, et propose surtout une visualisation permettant d'aisément voir à quel point deux personnalités apparaissent régulièrement ensemble ou non.

Yang et al. (2009) appliquent quant à eux cette problématique à la causalité entre événements, et sont donc plus proches dans l'esprit de la tâche proposée par Nallapati et al. (2004). Leur graphe reste néanmoins acyclique et organisé selon un axe temporel, deux caractéristiques pertinentes au vu du type de relation étudiée. Le graphe proposé par les auteurs est construit selon une similarité lexicale, et un seuil est utilisé afin de déterminer quels liens sont pertinents ou non (Yang et al., 2009). Illustré par les réactions d'un État, la Russie, aux attaques terroristes qu'il a subies (voir la figure 2.3), ce type de graphe permet de mieux comprendre les implications des événements successifs.

Une des applications les plus intéressantes de ces graphes est que l'on y trouve de

Event	Event Summary	Num. of doc.	Start time	End time
1	Chechen terrorists seized the Beslan school with hostages, negotiated, freed some hostages	5	2004-09-02 01:46	2004-09-03 07:08
2	Special task force assaulted terrorists and hundreds of hostages were dead	3	2004-09-03 14:46	2004-09-05 05:14
3	Responses of different parties on the Beslan school hostage tragic	5	2004-09-04 15:45	2004-09-07 13:04
4	Russia approached to identify the suspects of Beslan tragedy	6	2004-09-06 01:07	2004-09-08 17:54
5	Russia conducted investigation and determined to put terrorists on trial	4	2004-09-08 15:44	2004-09-24 11:36
6	Beslan school resumed classes after the hostage tragic	3	2004-09-14 08:12	2004-09-15 12:33
7	Russia claimed to strike Chechen terrorism	3	2004-09-14 08:52	2004-09-17 12:38
8	Russia's successive efforts against terrorism	3	2004-09-29 12:01	2004-12-17 13:53

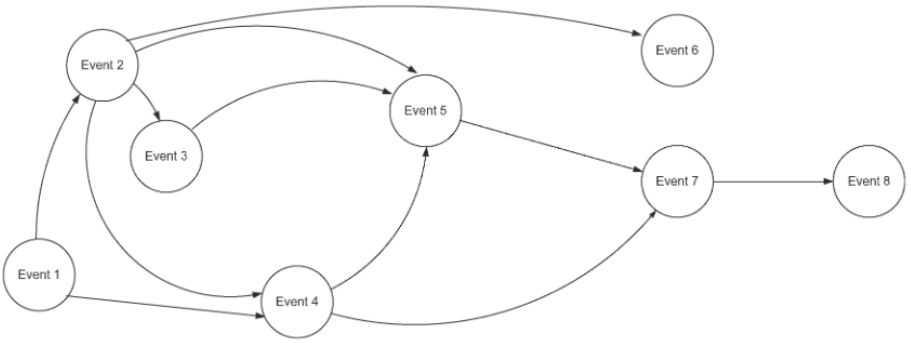


FIGURE 2.3 – Exemple d'organisation d'actualités sous forme de graphe acyclique (Yang et al., 2009).

façon explicite la notion de chemin entre deux documents. Dans le cas où le graphe ne disposerait que d'une seule composante connexe, on serait en mesure de trouver un chemin entre n'importe quelle paire d'articles. Cette idée a déjà été explorée par Shahaf et Guestrin (2010), sans toutefois s'appuyer sur de telles structures, mais en formulant cette tâche comme un problème d'optimisation dans lequel l'objectif est de construire une suite de documents reliant deux articles désignés arbitrairement par un utilisateur.

2.2.4 Hyperliage multimédia

L'hyperliage multimédia (ou *hyperlinking*) trouve ses origines dans la génération automatique de liens hypertextes (Wilkinson et Smeaton, 1999). Cette génération automatique peut avoir des objectifs multiples tels que la citation (référence à un autre hypertexte), l'organisation du contenu (*e.g.*, table des matières), la recommandation, … L'hyperliage correspond à son extension au multimédia, dans le but d'obtenir des hypermédias (Eskevich et al., 2013). Bien que les liens créés puissent être adaptés à chaque utilisateur par le biais d'outils de recommandation (Brusilovsky, 1998), la tâche d'hyperliage est généralement considérée comme indépendante de l'utilisateur et fondée sur le contenu des documents. Elle correspond davantage à un enrichissement du contenu qu'à une recommandation.

La mise en œuvre de l'hyperliage sur des collections vidéos mono-sources a été étudiée lors des campagnes d'évaluation de MediaEval (Eskevich et al., 2014), puis TRECVid (Over et al., 2015; Awad et al., 2016), et a d'abord été couplée à une problématique de détection de segments vidéos d'intérêt (*Search and Hyperlinking*) avant d'être envisagée de façon individuelle (*Hyperlinking*). Il s'agit, pour chaque segment d'intérêt (appelés ancres ou *anchors*), de proposer automatiquement une liste de segments cibles (ou *targets*) recommandés à l'utilisateur sans avoir accès à son besoin d'information spécifique. Étant donné la difficulté à caractériser de manière objective le bien-fondé de la création

d'un lien entre deux segments de vidéos, l'évaluation de l'hyperliage est réalisée au travers d'évaluations humaines dans lesquelles les annotateurs sont amenés à juger de la pertinence du lien (Eskevich et al., 2017).

Les différents systèmes ayant participé à ces campagnes d'évaluation se fondent en majorité sur les similarités lexicales des transcriptions des vidéos ainsi que sur les similarités visuelles (Guinaudeau et al., 2012a; Le et al., 2014; Cheng et al., 2015; Pang et Ngo, 2015), y ajoutant parfois des informations complémentaires issues des métadonnées (De Nies et al., 2013; Şimon et al., 2014). Si cette approche consistant à lier les segments les plus similaires possibles permet généralement d'obtenir de bons résultats lors de la campagne d'évaluation, elle résulte en une faible diversité des liens générés, comme nous le montrons dans la section 6.3.2. Cette problématique de la diversité dans le cadre de l'hyperliage est explorée plus largement dans le chapitre 6.

2.3 Systèmes complets

Les collections de documents journalistiques, audio, vidéos ou textuels, sont largement répandues et relativement faciles d'accès. De plus, elles sont quotidiennement manipulées par le grand public. Ces différents facteurs expliquent qu'elles soient très régulièrement utilisées comme base pour de nombreuses recherches dans les domaines du multimédia ou du traitement automatique des langues, telles que le résumé automatique (Hong et al., 2014), l'extraction d'entités nommées (Ratinov et Roth, 2009), la traduction automatique (Luong et al., 2015), la transcription automatique (Schlippe et al., 2013), ou l'enrichissement de données (Morang et al., 2005). Dès lors, il est naturel de vouloir réunir ces différentes fonctionnalités au sein d'un système complet, utilisable par le grand public ou les professionnels de l'information. Nous présentons dans cette section trois systèmes complets : Informedia (Hauptmann et Witbrock, 1997), l'un des premiers explorateurs d'actualités vidéo, Fischlàr News (Smeaton et al., 2001), similaire dans l'esprit mais plus récent et disposant de davantage de fonctionnalités, et FishWrap (Chesnais et al., 1995), un système multisources dédié à l'actualité au format texte, et qui proposait dès le milieu des années 90 des approches collaboratives permettant de mettre en avant les actualités jugées les plus importantes.

2.3.1 Informedia

D'abord conçu comme un outil dédié à l'éducation et au divertissement (Christel et al., 1994), Informedia a ensuite évolué afin de proposer à ses utilisateurs de consulter de larges archives vidéos dédiées à l'actualité (Hauptmann et Witbrock, 1997). Le processus d'adaptation de cet outil au domaine de l'information est intéressant à suivre en cela qu'il expose les particularités de ce domaine. Si les vidéos éducatives étaient relativement courtes et formaient chacune un tout cohérent, ce n'est pas le cas des journaux télévisés, qui sont la principale ressource du système. La demi-heure d'actualités quotidienne doit en effet être segmentée afin de séparer les divers sujets abordés dans le journal considéré. Une autre différence primordiale est le passage d'une collection statique, qui n'évolue pas, à une collection dynamique, réactualisée chaque jour. Ce processus doit évidemment être entièrement automatisé, et si les outils permettant cette automatisation sont aujourd'hui répandus, ce n'était pas le cas dans les années 90 (Brown et al., 1995).

S'il n'est pas le premier à exister, Informedia comporte toutes les briques élémentaires des systèmes d'exploration d'archives télévisuelles qui suivront : segmentation en su-

jets, segmentation en scènes, transcription automatique, indexation fondée sur les transcriptions, et moteur de recherche. Le système a continué à évoluer au cours des années, intégrant par exemple la détection de visages ou de textes présents dans les vidéos (Wactlar et al., 1996), des résumés automatiques (Wactlar, 1999) – aussi développés dans des systèmes comme ANSES (Pickering et al., 2003) – et l'extension à d'autres langues que l'anglais (Hauptmann et al., 1998). Néanmoins, il subit les faiblesses de certains de ses composants, peu développés à cette époque. Ainsi, la transcription automatique affiche un taux d'erreur (*Word Error Rate* ou WER) de 65 % sur un journal télévisé (Hauptmann et Witbrock, 1997). Cette transcription a été identifiée par les créateurs d'Informedia comme le composant critique, et si les outils de transcription sont de plus en plus efficaces, il semble que leur taux d'erreur reste trop élevé pour que les textes qu'ils génèrent soient affichés aux utilisateurs (Hauptmann, 2005).

2.3.2 Fischlàr News

Le système Físchlár News (Smeaton et al., 2001) vise à permettre à ses utilisateurs de consulter des archives d'actualités vidéos. Ces vidéos sont extraites d'une unique émission quotidienne d'information irlandaise – le journal de RTE1 – d'une durée moyenne de 30 minutes. Chaque jour, l'émission est automatiquement récupérée, traitée, et ajoutée à Físchlár News. Les traitements correspondent à une segmentation de la vidéo afin de séparer chacun des sujets abordés dans l'émission, puis à un ajout des segments à la base de données, avec plusieurs types de représentations (extraction des *frames*-clés (*keyframes*), sous-titrage, représentation à des fins de calculs de similarité, ...). Le système dispose de deux principales fonctions : la consultation des actualités d'une date donnée, et la recherche de vidéos par mots-clés. La première fonction correspond à l'affichage de l'ensemble des sujets traités à une date donnée, et permet une navigation facile afin de pouvoir visionner les parties de journal intéressant l'utilisateur. Ces parties de journal correspondent à l'ensemble des sujets abordés dans l'émission. Une fois un sujet sélectionné par l'utilisateur, une segmentation en plans de vue du sujet vidéo – chacun de ces plans étant associé à des résumés textuels – lui est proposée afin de lui permettre de sauter rapidement à une partie du reportage, ou, par exemple, de ne pas visionner l'introduction du journaliste en plateau. La seconde fonction permet, à l'aide d'un moteur de recherche à base de mots-clés, de retrouver l'ensemble des segments de journaux répondant à la requête, ordonnés de façon temporelle. Le système n'utilisant qu'une unique émission quotidienne, la quantité de sujets retournés par la requête reste relativement faible, et est donc tout à fait ergonomique pour les utilisateurs.

Une fonctionnalité de recommandation est également disponible. Ainsi, l'utilisateur peut évaluer les segments de son choix en indiquant s'ils l'intéressent ou non. Il recevra ensuite chaque jour un mail récapitulant les actualités de la veille qui sont susceptibles de l'intéresser. D'autres systèmes se sont également penchés sur la problématique de la recommandation d'actualités. C'est notamment le cas de NewsFlash, qui utilise les profils de ses utilisateurs, récupérés de façon implicite, afin d'améliorer leur moteur de recherche via une extension de requêtes profilée (Haggerty et al., 2003).

Fischlàr News se prête bien à une utilisation quotidienne, et permet à l'utilisateur de se tenir au courant de l'actualité en regardant les segments du journal télévisé qui l'intéressent. Néanmoins, il est probable que ce système ne puisse pas augmenter son nombre de sources sans sacrifier à sa facilité d'utilisation. Chaque journal TV comporte en effet une vingtaine de sujets (Smeaton et al., 2004), ce qui rend leur affichage exhaustif aisé, et garantit l'absence de redondance entre sujets.

2.3.3 FishWrap

FishWrap est un système conçu au MIT permettant notamment la consultation d'actualités issues de nombreuses sources telles que les journaux en ligne, mais également des actualités liées à la vie du campus (Chesnais et al., 1995). L'intérêt principal de ce système vient du fait qu'il ait été l'un des premiers à utiliser plusieurs sources d'actualités (Reuters, Associated Press, The Boston Globe, ...). Le standard RSS n'étant pas encore développé à cette époque, ces sources arrivaient sous différents formats (y compris mail), et étaient ensuite unifiées au sein d'une même structure. Les utilisateurs du système étaient amenés à répondre à trois questions lors du premier lancement de l'interface : leur ville d'origine (afin de leur fournir des actualités locales), leur lien avec le MIT (afin de recevoir les actualités du campus dédiées à leur groupe), et leurs intérêts (afin de pouvoir leur proposer des actualités adaptées à leurs préférences).

L'interface principale fournit aux utilisateurs une liste d'articles filtrés selon leurs préférences explicites (les réponses aux questions posées au lancement du logiciel), et implicites (articles visualisés plus tôt, popularité d'un article, ...). Une seconde fonctionnalité est la présence d'une *Page One*, qui correspond à un *best of* des articles récents selon les utilisateurs. En effet, les articles les plus vus sont automatiquement ajoutés à cette page, qui peut être considérée comme une visualisation des actualités les plus importantes.

La plupart des défis techniques résolus par ce système se trouvent aujourd'hui largement simplifiés par l'existence d'outils performants pour l'extraction et le formatage d'actualités. Néanmoins, FishWrap a été l'un des premiers systèmes à proposer l'utilisation d'une grande variété de sources, réunies au sein d'une même interface. Ce choix n'a été que peu ou pas repris dans les systèmes qui l'ont suivi, qui se concentrent généralement sur l'exploitation d'une unique source d'information.

Conclusion

Comme nous l'avons décrit dans ce chapitre, il existe de nombreuses façons d'organiser les collections d'actualités. Une première étape consiste invariablement à regrouper les articles discutant un même sujet. Des mesures de similarité fondées sur des représentations lexicales sont généralement suffisantes à ce stade. Lorsque l'on considère des structurations plus fines des collections, on remarque que la dimension temporelle de l'actualité occupe généralement une très grande place dans la littérature et est logiquement au centre de la plupart des approches. Ceci est cohérent avec l'importance de l'aspect temporel tel que décrit par les professionnels en section 1.2.3. Des structures arborescentes permettent de mieux rendre compte, au-delà de l'aspect temporel, des relations de causalité qui peuvent exister entre les événements. Les structures les plus abouties sont les graphes, qui peuvent représenter d'autres types de relations entre documents en s'intéressant aux personnages qui y figurent, aux liens de causalité qui peuvent exister entre les sujets abordés, ou d'autres encore.

Plusieurs systèmes complets permettant la consultation d'actualités ont été conçus, et disposent généralement de nombreux composants. Le résumé automatique y occupe une place importante, et est considéré comme l'un des meilleurs moyens de réduire la surcharge informationnelle. Ces systèmes sont la plupart du temps conçus pour une consultation quotidienne, lors de laquelle l'utilisateur souhaite connaître les informations récentes susceptibles de l'intéresser. Ils sont néanmoins peu adaptés à la navigation au sein

de grandes collections s'étalant sur plusieurs mois, et ne proposent que des outils classiques de recherche d'information pour consulter des documents plus anciens.

L'un des objectifs du projet LIMAH, discuté dans le chapitre suivant, ainsi que les principaux travaux de cette thèse, visent à répondre aux besoins des professionnels exprimés dans la section 1.2.3 et à améliorer la consultation d'actualités au moyen d'hypergraphes. L'objectif consiste, en combinant plusieurs types de liens entre documents tout en s'assurant des bonnes propriétés topologiques de l'hypergraphe construit, à faciliter l'exploration de grandes collections multisources et construites sur de longues périodes.

Chapitre 3

Le projet LIMAH

Introduction

Cette thèse s'est déroulée dans le cadre du projet CominLabs LIMAH (*LInking Media in Acceptable Hypergraphs*). Il s'agit d'un projet multidisciplinaire visant à répondre aux différentes problématiques introduites dans les deux chapitres précédents. Le projet vise notamment à enrichir des collections multimédias et multisources d'actualités en créant des liens explicites entre documents. Il s'intéresse également à la structuration de cours en ligne et à ses effets sur les apprenants. Enfin, il étudie les implications légales et l'acceptabilité des systèmes conçus lors du projet. Il réunit des équipes des laboratoires IRISA et LS2N (anciennement LINA) pour les domaines du multimédia et du traitement des langues, PREFICs et CRPCC pour l'étude des interfaces et leur acceptabilité, et IODE pour les aspects législatifs soulevés, notamment au niveau européen.

Dans ce chapitre, nous présentons en détail les objectifs du projet LIMAH ainsi que les partenaires associés. Nous décrivons ensuite la construction d'un corpus journalistique réalisée dans le cadre de cette thèse et visant à répondre à plusieurs des problématiques soulevées.

3.1 Enjeux et objectifs

Dans cette section, nous décrivons quatre thématiques de recherche abordées dans le projet LIMAH et la façon dont elles s'inscrivent dans un tout cohérent.

3.1.1 Construction d'hypergraphes navigables

L'un des objectifs du projet LIMAH consiste à organiser de façon pertinente de grandes collections multimédias d'actualités ou éducatives. Cette thèse en particulier s'intéresse aux contenus multimédias journalistiques multisources. En s'appuyant sur le principe d'hyperliage présenté dans la section 2.2.4 et sur la structuration en graphes discutée dans la section 2.2.3, le projet vise à construire et exploiter des hypergraphes de documents d'actualités qui peuvent être des articles de presse, des articles de blogs, des *podcasts* radio, des vidéos extraites de chaînes d'information, … Ces multiples formats

sont nécessairement issus de différentes sources, ce qui constitue un cas d'étude rarement approché par la recherche. En effet, de multiples événements sont discutés par différents médias, entraînant une redondance plus importante des informations disponibles. Les caractéristiques du corpus, dont la construction a été l'une des premières étapes de cette thèse, sont décrites plus précisément en section 3.2.

Les hypergraphes considérés sont à rapprocher de la notion d'hypermédias, version multimédia de l'hypertexte, un concept largement étudié dans le cadre du web. Ils se différencient de la structure classique de graphes d'actualités présentés dans la section 2.2.3 en cela qu'ils vont au-delà des considérations purement chronologiques ou causales développées jusqu'à présent. Les hypergraphes peuvent en effet incorporer différents types de liens, et permettre une navigation éclairée dans laquelle le type de relation existant entre deux documents est décrit de façon explicite. Ils sont une généralisation de la tâche d'hyperliage à l'ensemble d'une collection.

Ces structures, si elles sont construites de façon raisonnée, peuvent être aisément implémentées au sein d'un système complet tel que ceux décrits dans la section 2.3 et permettre une navigation directe de l'ensemble de la collection, dans une optique d'exploration plutôt que de recherche d'information. Elles peuvent également servir de base à des traitements annexes, tels que le résumé multidocuments. Ainsi, les documents fortement connectés les uns aux autres, qui forment un groupe cohérent, peuvent faire l'objet d'un résumé automatique afin de limiter la surcharge informationnelle qui existe lorsque de grandes collections sont en jeu.

3.1.2 Segmentation et structuration de vidéos éducatives

L'actualité n'est pas le seul domaine à pouvoir bénéficier d'une structuration plus fine des collections. Bien que les cours en ligne existent depuis longtemps à travers des formations à distance, le domaine éducatif a récemment connu de nombreux changements avec le succès grandissant des cours en ligne ouverts et massifs (*Massive Open Online Courses* ou MOOC), qui permettent à des centaines de milliers d'internautes de suivre des cours dispensés par des experts en s'émancipant des contraintes de localisation ou des coûts d'inscription. Cette évolution a encouragé l'utilisation de nouveaux outils pédagogiques, renforçant la place du multimédia, et notamment des supports vidéos.

Il est connu depuis longtemps que la structuration pertinente, par exemple par le chapitrage, des cours dits « classiques » au format écrit est un paramètre déterminant dans l'efficacité des apprenants. Elle est notamment utile dans l'apprentissage, la recherche d'informations, ou la révision. Le projet LIMAH vise donc à étudier comment la segmentation, et davantage encore la structuration, de vidéos éducatives peuvent permettre d'améliorer ces aspects. Les apprenants eux-mêmes se trouvent au cœur de cette étude, qui a davantage pour objet de comprendre leurs besoins qu'à construire automatiquement les segmentations et structures jugées pertinentes.

3.1.3 Analyse d'opinion et contenus utilisateurs

La grande diversité des documents d'actualités considérés pousse à l'analyse fine des opinions qu'ils transmettent. Bien que les articles des médias de référence soient généralement dépourvus de marqueurs d'opinion, ce n'est pas le cas des blogs, ni des autres médias semi-professionnels ou encore des réseaux sociaux. Le projet LIMAH vise à améliorer les techniques d'analyse de ces opinions, à la fois dans les communications

semi-professionnelles telles que les blogs, mais également dans celles qui proviennent des consommateurs de l'actualité, et qui disposent de plus en plus d'endroits dans lesquels s'exprimer (sections commentaires des articles, réseaux sociaux, ...). Or, ces communications externes mettent en avant de nouvelles problématiques : elles sont plus régulièrement dégradées (abréviations, erreurs grammaticales ou lexicographiques, ...), font appel à davantage de figures de style et de modes (parodie, humour, second degré, métaphores, ...), font souvent référence de manière implicite au document ou au média qu'elles commentent, voire à des notions extérieures. Le projet LIMAH étudie donc non seulement les moyens de composer avec ces nouvelles difficultés, mais analyse également plus finement les cibles de l'opinion ainsi que l'évolution des opinions dans le temps.

3.1.4 Droit des données et des enrichissements

Les articles de presse publiés ainsi que les vidéos d'actualités sont soumises à une législation disposant d'un long historique. Il s'agit notamment de la propriété intellectuelle, des droits d'auteurs et de ceux des ayants-droits. Néanmoins, un flou juridique existe sur les données qui y sont associées, et particulièrement celles qui impliquent la génération de nouveaux contenus. Un système de résumé automatique permettant de condenser une dizaine d'articles en une centaine de caractères fait-il une utilisation illégale des articles originaux s'il publie les résumés générés ? Sachant que ces outils s'appuient la plupart du temps sur l'extraction sans modification des phrases les plus importantes des articles, on pourrait considérer que si la fraction des phrases extraites d'un même article est suffisamment importante, alors le système contrevient aux droits d'auteurs.

De la même façon, la création de liens entre documents peut-elle être illégale ? Plusieurs jugements semblent pour l'instant contradictoires à ce sujet. Par exemple, la création d'un lien vers un contenu protégé par le droit d'auteur est illégale si la personne ayant créé ce lien l'a fait en ayant connaissance de l'illégalité de la mise à disposition de l'oeuvre [1] (qui aurait pu être mise en ligne de façon légale, avec l'accord de l'auteur, par un tiers). Un exemple facilitant quant à lui la création libre d'hyperliens est donné avec le jugement de la cour de Cassation du 31 mars 2016 [2]. Dans ce jugement, qui vise une personne ayant relayé sur son site une vidéo de menaces de mort sans en avoir été l'auteur, la Cour explique que « l'article 433-3 du code pénal n'incrimine pas le fait de faciliter la diffusion de menaces de mort ». Ce sont ces questions de droits d'auteurs et de droits de lier, illustrées par les exemples précédents, qui sont explorées dans le cadre du projet.

Toutes les différentes problématiques décrites dans cette section doivent être instanciées afin de disposer d'une base de réflexion et d'étude. Dans ce sens, nous proposons la construction d'un corpus pour répondre à une partie d'entre elles.

3.2 Corpus : construction et caractéristiques

Il n'existe pas, à notre connaissance, de corpus d'actualités multisources et multimédia, en langue française, reflétant la diversité des sources disponibles à l'heure actuelle. Nous avons donc, dans notre thèse, proposé la construction d'un nouveau corpus prenant en compte aussi bien des documents issus de la presse en ligne établie (Le Monde, Le Figaro, Libération, ...), d'émissions télévisuelles disponibles en *replay* en ligne ou de

1. Arrêt C-160/15 de la Cour Européenne www.curia.europa.eu/juris/documents.jsf?num=C-160/15
2. www.legalis.net/jurisprudences/cour-de-cassation-chambre-criminelle-arret-du-31-mars-2016/

podcasts radiophoniques. Nous y avons également associé les données générées par le grand public qui consomme ces informations, notamment au travers des sections commentaires des différents sites, ainsi que des réseaux sociaux.

Dans cette section, nous décrivons le corpus construit afin de répondre aux problématiques de structuration des collections d'actualités, ainsi que d'étude des sentiments et des opinions produits par ceux qui consomment l'actualité.

3.2.1 Objectifs et composition

Nous avons construit un corpus se composant de documents issus de sources journalistiques. Son but est de regrouper des documents de modalités différentes (audio, vidéo, écrit) ainsi que des types de discours différents (articles de fond, brèves, blogs, tweets, interviews, ...). Certains documents sont également accompagnés des commentaires laissés par les utilisateurs (commentaires d'article de presse, tweets, ...). Tous les documents ont été récupérés entre le 20 mai 2015 et le 8 juin 2015, via des flux RSS. Les documents complets, qu'ils soient au format HTML, dans un format audio, dans un format vidéo, ou dans un autre format (*e.g.*, json pour Twitter) ont été stockés. La construction de cette ressource a été réalisée avec l'aide de Sébastien Campion et de Grégoire Jadi. Le code source ayant servi à sa construction est disponible en ligne [3].

Nous décrivons ici les différents documents qui composent le corpus construit. Nous présentons successivement les statistiques du corpus modalité par modalité, en commençant par les documents web (articles de presse et blogs), les documents audio (*podcasts* radio), les documents vidéo (émissions et journaux télévisés), et enfin les données issues de réseaux sociaux. Pour chaque section, des informations statistiques sont fournies. La table 3.1 indique le nombre de documents présents pour chaque catégorie.

Type	Nombre de documents
Presse	4 966
Radio	1 556
Video	290

TABLE 3.1 – Nombre de documents par type.

3.2.2 Documents web.

La première catégorie de documents récupérés comprend les pages webs. Ces pages incluent les articles de journaux et articles de blogs. La liste des sites utilisés sont indiqués dans la table 3.2. Si les flux RSS sont une porte d'entrée intéressante, il est nécessaire de récupérer les pages web entièrement afin de bénéficier des commentaires des utilisateurs, de la mise en page, et de métadonnées supplémentaires. Les billets de blog permettent d'exposer davantage de déclarations d'opinions et proposent parfois de mettre en lumière certains aspects de l'actualité.

Plusieurs données ont été extraites à partir du HTLML des pages web obtenues :

3. www.github.com/sildar/limah

Source	Type de la source	Nombre de documents
Le Monde	Presse	1 802
Le Point	Presse	1 029
Le Figaro	Presse	812
Libération	Presse	683
Huffington Post	Presse	640
Blogs le Monde	Blog	137
Blogs le Figaro	Blog	23

TABLE 3.2 – Documents web.

— titre ;
— texte (contenu principal) ;
— HTML du contenu principal ;
— date de publication ;
— URL ;
— source (*e.g.*, Le Monde) ;
— catégorie (blog ou presse) ;
— image d'illustration ;
— auteur (nom du journaliste quand présent) ;
— description (texte d'introduction quand présent).

Ces données ont été acquises grâce à la bibliothèque Python Newspaper [4]. Lorsque cet outil s'est révelé insuffisant pour obtenir un contenu principal propre, des scripts d'extraction spécifiques ont été développés avec la bibliothèque Scrapy [5]. Plusieurs traitements linguistiques ont été ensuite été réalisés sur le contenu principal :

— découpage en phrases et en mots ;
— racinisation (*stemming*) et étiquetage morpho-syntaxique (*part-of-speech tagging*) ;
— entités nommées (personnes, lieux, ...) ;
— repérage des liens hypertextes présents dans le texte de l'article ;
— extraction dee mots-clés.

Le découpage en phrases et en mots, ainsi que la racinisation et l'étiquetage morpho-syntaxique ont été réalisés via l'outil Apache OpenNLP [6]. La bibliothèque Newspaper a été utilisée pour l'extraction d'entités nommées et de liens hypertextes.

Source	Nombre de phrases (moy)	Nombre de mots (moy)
Le Monde	21.7	651
Le Point	19.6	592
Le Figaro	7.4	259
Libération	28.4	792
Huffington Post	21.0	615
Blogs le Monde	30.6	884
Blogs le Figaro	26.8	611
Moyenne pondérée	20.1	597
Médiane	15	477

TABLE 3.3 – Statistiques sur les documents web.

Quelques statistiques supplémentaires ont été calculées et sont présentées dans la table 3.3. La figure 3.2 donne une représentation visuelle de ces statistiques. On remarque une grande diversité dans la taille des documents, avec des articles beaucoup plus courts

4. www.github.com/codelucas/newspaper
5. www.github.com/scrapy/scrapy
6. www.opennlp.apache.org

pour le Figaro. Un exemple d'article du Figaro est donné en figure 3.1. Le format de l'article est recréé grâce aux différentes métadonnées récupérées, ainsi que grâce au HTML du contenu principal. Les articles de blogs sont en moyenne plus longs que les articles classiques du journal qui les publie.

FIFA: Vladimir Poutine félicite Sepp Blatter

Le président russe Vladimir Poutine a adressé un télégramme de félicitations au président de la Fédération internationale de football (Fifa), Sepp Blatter, pour le féliciter à l'occasion de sa réélection.

"Le chef de l'État russe a dit son espoir que l'expérience, le professionnalisme et la haute autorité dont il jouit aideront (Joseph) Blatter à l'avenir à encourager le développement du football à travers le monde", a déclaré le Kremlin dans un communiqué. La Russie souhaite coopérer avec la Fifa de manière générale, et tout particulièrement en vue de préparer la phase finale de la Coupe du monde 2018, qu'elle organisera.

Blatter a été réélu vendredi président de la Fifa après le retrait de son rival, le prince jordanien Ali ben al Hussein, du second tour de scrutin auquel le patron du football mondial avait été contraint. Ignorant les appels à la démission qui s'étaient multipliés ces derniers jours après les scandales de corruption à l'échelle planétaire affectant l'organisation, Joseph "Sepp" Blatter, 79 ans, a tenu tête et obtenu le droit d'assurer un cinquième mandat de quatre ans au terme duquel il a promis de laisser "une Fifa plus forte" à son successeur.

LIRE AUSSI :
» Blatter : «Je pardonne à tout le monde mais je n'oublie pas»
» Blatter réélu à la tête de la Fifa

FIGURE 3.1 – Un exemple d'article du Figaro.

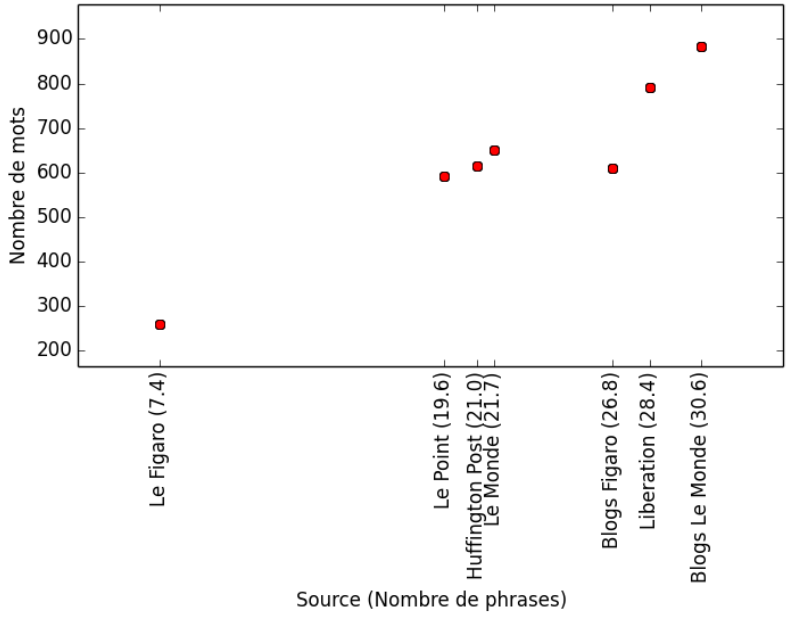


FIGURE 3.2 – Rapport entre le nombre de mots et le nombre de phrases.

L'extraction automatique de mots-clés est un bon indicateur pour connaître les principales thématiques abordées dans un corpus. On utilise ici le *Term Frequency-Inverse Document Frequency* (tf-idf) afin d'ordonner les mots les plus importants dans chaque document. L'idf est calculé sur l'ensemble du corpus. La table 3.4 donne les vingt mots-clés les plus fréquents dans les documents de presse, ainsi que le nombre de documents dans lesquels ils apparaissent.

Comme le montre la table 3.4, les thématiques les plus abordées sont le scandale de la FIFA (fifa, blatter, président, football), le renommage du parti de l'UMP en Les Républicains (parti, sarkozy, nicolas, républicains), le tournoi de Roland Garros (roland, garros), les flux migratoires (migrants), et les difficultés économiques de la Grèce (ministre, grèce, milliards, dollars).

Mot-clé	Fréquence
parti	141
fifa	130
sarkozy	115
roland	90
ministre	90
garros	86
groupe	82
grèce	81
milliards	81

TABLE 3.4 – Mots-clés les plus fréquents dans les documents de presse.

3.2.3 Documents audio

Les émissions de radio d'information ainsi que des chroniques traitant de l'actualité sont ciblées. Certaines émissions impliquent plusieurs orateurs en même temps (*e.g.*, Les Grandes Gueules), tandis que d'autres se concentrent sur un présentateur (*e.g.*, Journal de 13h) ou des interviews de personnalités ou d'auditeurs.

Les *podcasts* récupérés sont décrits dans la table 3.5. Chaque document a été transcrit automatiquement et segmenté thématiquement automatiquement à l'aide de l'approche développée par Guinaudeau et al. (2012b). Le nombre de documents apparaissant dans la table 3.5 représente ce nombre de segments, et non le nombre d'émissions distinctes.

Émission	Source	Genre	Nombre de documents
Bourdin Direct	BFM	Public	556
Chroniques	France Inter	Chronique	392
Les Grandes Gueules	BFM	Public	283
Divers	France Inter	Emission	144
Carrément Brunet	BFM	Public	50
Journal de 8h	France Culture	Journal	67
Journal de 13h	France Inter	Journal	46

TABLE 3.5 – Documents audio.

Les métadonnées extraites pour chacun de ces documents sont les suivantes :
— nom de l'émission ;
— transcription automatique ;
— segmentation automatique ;
— date de publication ;
— URL du *podcast* ;
— source (e.g. RMC).

Les traitements linguistiques supplémentaires suivants ont été réalisés :

— découpage en pseudo-phrases et en mots ;
— racinisation et étiquetage morpho-syntaxique ;
— détection des entités nommées (personnes, lieux, ...) ;
— extraction de mots-clés.

La table 3.6 fournit quelques statistiques sur la taille des différents segments.

Comme pour la presse, les mots-clés ont été extraits automatiquement des transcriptions des documents radio. La table 3.7 donne les vingt mots-clés les plus fréquents ainsi que le nombre de documents dans lesquels ils apparaissent.

Émission	Nombre de phrases (moy)	Nombre de mots (moy)
Bourdin Direct	55.7	835.2
Chroniques	36.8	516.6
Les Grandes Gueules	72.9	1 154.6
Divers	49.5	786.5
Carrément Brunet	115.1	1 446.9
Journal de 8h	20.6	329.7
Journal de 13h	31.1	407.2
Moyenne	52.9	789.7
Médiane	49.5	697.4

TABLE 3.6 – Statistiques sur les documents audio.

Mot-clé	Fréquence
euh	255
heures	109
gens	91
hui	76
parti	68
merci	60
fifa	60
france	59

TABLE 3.7 – Mots-clés les plus fréquents à la radio.

On peut s'étonner de la présence du terme « euh » en première position des mots-clés. En effet, « euh » étant fréquemment attendu, son *Inverse Document Frequency* (idf) devrait être très faible et son tf-idf devrait donc être réduit. En fait, cet idf est calculé sur l'ensemble du corpus, y compris sur les documents de presse écrite, qui ne comportent logiquement aucune fois le terme « euh ». De plus, ce terme est très peu fréquent dans des émissions de radio de type journal d'information, et au contraire extrêmement fréquent lors d'émissions faisant intervenir des auditeurs (*e.g.*, Bourdin Direct). Il apparaît donc comme fortement discriminant pour certains documents, et se retrouve donc en tête de liste. Le même raisonnement peut être appliqué à d'autres termes tels que « heures » (« Il est huit heures »), « merci », « rmc », ... Il apparaît que les mots-clés extraits des documents audio sont moins exploitables du fait de ce bruit. On repère tout de même des références aux thématiques abordées dans la presse (FIFA, Les Républicains).

3.2.4 Documents vidéos

Des émissions quotidiennes ou hebdomadaires ont été récupérées, parmi lesquelles des journaux télévisés, des débats politiques, des magazines d'information. La liste des sources utilisées et des quantités récupérées est disponible dans la table 3.8. Comme pour les radios, le nombre de documents correspond au nombre de segments après segmentation automatique. D'autres émissions ont été récupérées mais n'ont pas été incluses dans le corpus. En effet, celles-ci étaient très irrégulières (*e.g.*, les émissions hebdomadaires) et/ou ont été récupérées de façon incomplète.

Les métadonnées extraites pour chacun de ces documents sont les suivantes :

Émission	Source	Type	Nombre de Documents
JT 20h	France 2	Journal	160
JT 13H	France 2	Journal	41
C dans l'air	France 5	Actu Débats	37
C a vous	France 5	Actu Divertissement	35
JT 06H	France 2	Journal	17

TABLE 3.8 – Documents vidéos.

— nom de l'émission ;
— transcription automatique ;
— segmentation automatique ;
— date de publication ;
— URL de la vidéo ;
— source (*e.g.*, France 2).

Les informations supplémentaires suivantes ont été extraites :

— découpage en pseudo-phrases et en mots ;
— racinisation et étiquetage morpho-syntaxique ;
— détection des entités nommées (personnes, lieux, ...) ;
— extraction de mots-clés.

Émission	Nombre de phrases	Nombre de mots
JT 20h	66.3	628
JT 13H	50.9	515
C dans l'air	108.2	1 230
C a vous	49	492.5
JT 06H	11.2	127
Moyenne	64	643
Médiane	51	515

TABLE 3.9 – Statistiques sur les documents vidéos.

La table 3.9 récapitule les différentes tailles des documents vidéos. On peut noter que le rapport entre nombre de phrases et nombre de mots est constant. Cela peut être attribué à deux facteurs : les émissions traitées sont majoritairement très préparées et avec prompteur (*e.g.*, journaux télévisés), et évitent donc les phrases très longues régulièrement présentes dans des émissions de type radio. Seule l'émission *C dans l'air* possède un nombre de mots par phrase plus important, probablement dû au fait qu'elle consiste à poser des questions à des experts qui ne sont pas soumis à un prompteur.

Une fois encore, les mots-clés ont été extraits automatiquement des documents vidéo. La table 3.10 donne les vingt mots-clés les plus fréquents ainsi que le nombre de documents dans lesquels ils apparaissent.

Tout comme pour la radio, le terme « euh » est discriminant pour certaines émissions. Il est ainsi rare dans les journaux télévisés, mais très présent dans les émissions plus longues fondées sur des questions-réponses telles que *C dans l'air*. Certaines thématiques réapparaissent telles que le scandale de la FIFA, et de nouvelles émergent telles une mention des « élèves » et de la « réforme » qui correspondent à la réforme de l'enseignement au collège et à la consultation des enseignants sur les nouveaux programmes qui ont eu lieu en mai 2015.

Mot-clé	Fréquence
euh	45
ans	16
france	12
soir	12
ville	10
gens	10
coalition	9
élèves	8
mois	8
fifa	8

TABLE 3.10 – Mots-clés les plus fréquents à la télévision.

3.2.5 Réseaux sociaux et commentaires utilisateurs

Les réseaux sociaux sont un lieu de prolongement de l'information. Les consommateurs de médias y échangent autour de l'actualité, dans des formes très diversifiées. Nous avons rassemblé tous les tweets publiés sur les comptes de journaux présents dans le corpus, ainsi que tous les tweets faisant mention d'un journal sur la période visée. Ces milliers de tweets ont ensuite été filtrés pour ne conserver que ceux qui font une mention explicite (via leur URL) d'un article de presse faisant partie de notre corpus. 15 940 tweets ont été obtenus de cette façon. Les mots-clés utilisés pour récupérer ces tweets sont indiqués dans la table 3.11. Les tweets ayant une taille limitée (140 caractères), on obtient sans surprise uniquement 1,3 phrase par tweet en moyenne.

La plupart des documents web obtenus disposent d'une section « commentaires ». Ces derniers peuvent également être récupérés et utilisés à des fins d'analyses (voir section 3.1.3). Il arrive que les commentaires répondant à un même article soient répartis sur plusieurs pages distinctes. Dans ce cas, l'ensemble des pages concernées a été capté. Néanmoins, étant donné que les pages ont été obtenues à des horaires fixes (durant la nuit), il est possible que d'autres commentaires soient apparus dans les heures ou jours suivant la récupération. Ces commentaires, déposés après notre passage, ne sont donc pas disponibles dans le corpus.

Mots-Clés
lemonde
lepoint
huffingtonpost
rmc
radiofrance
francetv
arte
lcp
afp

TABLE 3.11 – Mots-clés utilisés via l'API Twitter afin de récupérer les commentaires visant des documents journalistiques.

Conclusion

Le projet LIMAH est un projet multidisciplinaire visant à améliorer d'une part les outils de consultation d'actualités et, d'autre part, les outils éducatifs multimédias. Ces deux objectifs s'accompagnent de plusieurs thématiques connexes, telles que l'analyse précise d'opinion et de sentiment, mais également d'enjeux législatifs et ergonomiques. Afin de mettre en œuvre une partie de ces objectifs, nous avons créé un corpus multimédia composé d'articles de presse, de *podcasts* radios, et de vidéos d'actualités, qui sont représentatifs d'une large partie du paysage médiatique tel qu'il est proposé au grand public. Sa particularité, en dehors de ses aspects multimodaux, est qu'il est grandement multisources, et fait donc apparaître de nouvelles problématiques rarement traitées dans la recherche, telles que la détection de redondance entre les sources.

Avec ce corpus à disposition, cette thèse s'attache à construire une structuration pertinente des documents qui le composent ayant pour but d'aider grand public et professionnels à parcourir l'actualité dans les meilleures conditions possibles afin d'améliorer pour les premiers l'accès à une information multisources, et, pour les seconds, d'accéder efficacement à l'ensemble des informations publiées sur un sujet précis.

Deuxième partie

Construction d'hypergraphes navigables pour l'exploration d'actualités

Chapitre 4

Hypergraphes explorables

Introduction

L'utilisation de graphes pour la navigation au sein d'actualités permet de répondre à de nombreux enjeux liés à la consultation d'actualités, et notamment au besoin ressenti par le grand public de donner du sens à l'information. Néanmoins, les choix réalisés lors de la construction de ces graphes peuvent mener à des structures diamétralement opposées. On peut ainsi envisager des graphes aux multiples composantes, dans lesquels de petits groupes d'articles discutant une série d'événements sont fortement liés entre eux et déconnectés du reste de la collection. Ces grappes d'articles peuvent être assimilées aux *clusters* des agrégateurs, à ceci près que l'utilisateur est encouragé à consulter plusieurs éléments d'un même *cluster*. Il s'agit de l'approche la plus communément employée, et décrite dans la section 2.2.3.

Une autre approche, qui est celle défendue dans cette thèse, consiste au contraire à construire un graphe doté d'une unique composante, dans lequel, en suivant les liens suggérés pour l'ensemble des documents, on peut *in fine* atteindre chaque élément de la collection. L'utilisateur est alors libre de naviguer dans ce graphe, allant de document en document, et peut dériver de sujets en sujets. Cette approche induit cependant un risque, celui de perdre l'utilisateur en lui proposant un trop grand nombre de liens à suivre. Il devient alors capital de maîtriser la topologie du graphe, c'est-à-dire, entre autres, le nombre moyen de liens proposés pour chaque document, l'absence de cycles bloquants la navigation, ou encore la capacité à atteindre tout document.

Dans ce chapitre, nous commençons par définir précisément la structure choisie, à savoir l'hypergraphe. Nous détaillons notamment ses liens avec la tâche *d'hyperlinking*, et ses différences avec l'objet mathématique homonyme. Nous introduisons ensuite les traits souhaitables de cet hypergraphe en terme d'explorabilité, c'est-à-dire les caractéristiques topologiques qui le rendent exploitable pour une utilisation par des professionnels et le grand public. Enfin, nous différencions la notion d'explorabilité d'un graphe de celle de navigabilité.

4.1 L'hypergraphe, une structuration de données pensée pour la navigation

L'hypergraphe correspond à un graphe d'hypermédias. Au sein de cette structure, on peut naviguer de document en document, en suivant les liens construits. Dans cette section, nous donnons une définition mathématique de l'hypergraphe et mettons en avant ses différences avec la recommandation et la recherche d'information.

4.1.1 Définition de l'hypergraphe

L'hypergraphe dont nous parlons dans cette thèse correspond au graphe obtenu par un processus *d'hyperlinking*, c'est-à-dire de création de liens entre hyperdocuments, telle que décrite en section 2.2.4. Il est à distinguer de la structure mathématique d'hypergraphe, qui correspond à une généralisation des graphes dans laquelle un ensemble de nœuds peuvent être liés à plusieurs autres par un unique arc (Gallo et al., 1993). L'hypergraphe mathématique étant une généralisation des graphes, nos structures s'y apparentent bien évidemment, mais les spécificités de ce modèle ne sont pas employées.

Formellement, notre hypergraphe peut être défini de la même manière que les graphes décrits dans la section 2.2.3 : un ensemble de nœuds V, qui correspondent aux documents présents dans la collection, et un ensemble d'arcs E qui connectent les nœuds deux à deux, tel que : $\forall e \in E,\ e = (v_i, v_j)\ v_i, v_j \in V^2$.

Il reste donc à définir la sémantique portée par les nœuds (documents) et arcs (liens) de notre structure. Intéressons nous d'abord aux nœuds. Dans notre cas, les collections considérées sont multisources et multimédias. On y trouve donc des articles de presse, des journaux télévisés, des émissions de radio, et autres. S'il paraîtrait logique de conserver dans leur entièreté des articles de presse en ligne, souvent succincts et concentrés sur la description d'un unique événement, il paraîtrait également logique de segmenter les journaux télévisés ou radiophoniques afin de séparer, *a minima*, les sujets qui y sont abordés. Dans cette thèse, nous choisissons une segmentation en *informations*, plutôt qu'une segmentation en *événements*, bien que ce second concept soit plus souvent utilisé dans la littérature. La distinction entre information et événement est décrite ainsi par Neuveu et Quéré (1996) :

> *Tout d'abord, toute nouvelle n'est pas nécessairement un événement. Certes une information est le plus souvent un événement (au sens ordinaire de « quelque chose qui est arrivé ») porté à la connaissance d'un individu ou d'un public. Mais elle peut tout aussi bien concerner une situation, un état de choses ou les actions d'une personne, d'un groupe - les faits et gestes des détenteurs du pouvoir politique, par exemple. Ensuite, une information peut « faire événement » sans qu'elle relate un événement à proprement parler : par exemple, un projet de réforme présenté par un gouvernement peut retenir l'attention publique et être doté d'une signification ou d'une valeur qui le sort de l'ordinaire - sans qu'il s'agisse à proprement parler d'un événement (c'est-à-dire d'une occurrence singulière, imprévue, non répétable) : c'est plutôt un fait notable. Or un fait n'est pas un événement, au sens propre du terme.*

Cette définition de l'information nous semble donc mieux adaptée à notre cas d'étude que la notion d'événement. Nous y ajoutons la notion de portée argumentative de l'information, c'est-à-dire que les documents d'actualités n'ont pas uniquement un but infor-

matif, mais portent intrinsèquement un point de vue sur les faits qu'ils rapportent, et ce qu'il s'agisse d'éditoriaux ou de blogs, mais également de presse classique ou journaux télévisés (Emediato, 2011). Nous choisissons donc de ne pas masquer ces aspects informatifs et argumentatifs des articles, et relions des documents présentant des informations plutôt que des documents décrivant des événements.

En ce qui concerne la sémantique des liens, nous reprenons la définition de l'hyperliage telle que définie dans les tâches de MediaEval et TRECVid : il s'agit de suggestions indépendantes des profils des utilisateurs, fondées sur le contenu, visant à leur apporter davantage d'informations sur le document visionné, sans répondre directement à un besoin d'information explicite (Eskevich et al., 2017). Cette définition reste néanmoins assez floue, à dessein. En effet, le bien-fondé de la création d'un lien entre deux articles de presse relève souvent du subjectif (Ge et al., 2010). Dans le cas d'articles de presse dénonçant la corruption supposée de personnalités de premier plan, comme c'est le cas dans le corpus LIMAH avec l'affaire FIFA (voir section 3.2), créer des liens avec d'autres articles (par exemple l'attribution des coupes du monde de football à la Russie en 2018 et au Qatar en 2022) pourrait être interprété comme des soupçons supplémentaires de corruption, quand bien même les articles liés seraient neutres. Or, ces liens, non évidents, subjectifs et discutables sont précisément parmi les plus intéressants pour les professionnels et le grand public. Face à ce constat, la meilleure possibilité d'évaluation de la pertinence des liens créés entre deux documents semble être le recours à l'annotation humaine. C'est notamment l'approche utilisée dans le cadre de TRECVid (Eskevich et al., 2017). Dans le cadre de cette thèse, au vu des aspects subjectifs intrinsèques à l'hyperliage, nous nous contenterons de cette définition intuitive de la sémantique des liens entre documents.

Les hypergraphes peuvent servir à diverses fins. Dans le cadre de cette thèse, nous nous intéressons à ces structures car elles sont un moyen efficace de représenter et de matérialiser les liens entre informations, et sont donc des supports pertinents pour l'exploration de collections d'actualités. Ils peuvent également servir entre autres au regroupement spectral (*spectral clustering*) (Ding et al., 2001), à la détection de thématiques (Ertöz et al., 2004), ou à la génération de chaînes d'actualités (Shahaf et Guestrin, 2010).

4.1.2 Différences avec les moteurs de recherche et la recommandation

Les moteurs de recherche sont des outils destinés à répondre à un besoin d'information émanant d'un utilisateur. De nombreux enjeux sont associés à cette problématique, tels que les moyens de formulation de ce besoin et ses problèmes associés (ambiguïté, formulations différentes d'une même idée, ...) (Krovetz, 1997), la diversité des résultats fournis (Wang et al., 2006; Foster et Ford, 2003), ou encore l'adaptation à des profils utilisateurs (Speretta et Gauch, 2005). Ce besoin d'information, s'il n'est pas toujours clairement formulé, est précisément le problème auquel les moteurs de recherche tentent de répondre. Dans notre cas, nous ne cherchons pas à répondre à un besoin d'information précis, mais à permettre l'exploration d'une grande collection. Si le point de départ de cette exploration peut initialement être issu d'une recherche d'information de la part de l'utilisateur, la structure que nous cherchons à construire lui permettra de s'éloigner de son sujet d'intérêt premier, dans une optique de découverte et d'exploration d'un ensemble plutôt que de recherche d'un point précis. En ce sens, nous nous rapprochons davantage d'une approche de cueillette (*berry-picking*) (Bates, 1989) que de recherche d'information.

La recommandation, elle, se concentre sur le profilage des utilisateurs afin de proposer la consultation de documents susceptibles de les intéresser. Elle se divise en deux ca-

tégories (Balabanović et Shoham, 1997) : la recommandation fondée sur le contenu à travers laquelle un utilisateur se voit proposer des documents similaires à ceux qu'il a déjà consultés, et la recommandation collaborative dans laquelle un utilisateur se voit proposer des documents fréquemment consultés par d'autres utilisateurs lui ressemblant. Certaines problématiques de ce domaine sont communes avec la recherche d'information, comme la recherche de diversité dans les recommandations (Ge et al., 2010). Ces approches permettent aux utilisateurs une personnalisation du contenu qui leur est présenté, et sont particulièrement pertinentes dans le domaine de l'actualité (Bucy, 2004; IJntema et al., 2010). Dans notre cas, bien que les liens que nous proposons puissent être vus comme des recommandations de documents à lire, la structure que nous cherchons à créer se veut indépendante de l'utilisateur, et vise à ne pas filtrer l'information qui lui est présentée.

4.2 Navigabilité et explorabilité : les caractéristiques souhaitables d'un hypergraphe

L'hypergraphe, appliqué à toute une collection, permet l'exploration de cette collection. Dans une telle structure, l'utilisateur peut naviguer de document en document, jusqu'à potentiellement s'écarter totalement du sujet abordé par le premier document visité. Dans cette section, nous introduisons le concept d'explorabilité du graphe. Nous le différencions de la navigabilité et proposons un ensemble de propriétés nécessaires à la bonne explorabilité d'un hypergraphe.

4.2.1 Explorabilité

L'explorabilité, telle que nous la définissons dans cette thèse, correspond à la facilité de naviguer au sein d'une collection, ici représentée sous forme de graphe. Elle correspond à un critère d'utilisabilité et est similaire au concept de navigabilité des sites web (Wojdynski et Kalyanaraman, 2016), c'est-à-dire la capacité de trouver aisément l'ensemble des informations qu'ils contiennent. Dans le cas des sites web, cela se fait généralement par la mise à disposition de menus de navigation, qui dirigent vers différentes sections du site, elles-mêmes disposant généralement de sommaires, sous-sections, ... Ainsi, pour un site de presse, la première page reprend invariablement les actualités les plus importantes du moment. Des sections thématiques spécifiques (sports, international, politique, ...) sont accessibles en un clic, et de là, des sous-catégories, une chronologie des articles publiés, une mise en avant des documents les plus populaires, etc.

Dans le cadre des graphes, cela correspond également à la capacité d'accéder aisément à l'ensemble des informations disponibles. Cette notion a été partiellement abordée dans des travaux tels que les graphes de recommandation (Lamprecht et al., 2016), notamment dans le domaine musical où Seyerlehner et al. (2009) tentent d'améliorer l'explorabilité (sous la notion de *browsability* dans leurs travaux) des recommandations musicales. Dans leurs travaux, cette explorabilité se limite à la possibilité de consulter l'ensemble de la collection en allant de recommandation en recommandation, sans toutefois s'intéresser à la réelle accessibilité des documents, dont certains peuvent se trouver en bout d'une longue chaîne de recommandations successives. D'un point de vue mathématique, il s'agit de créer un graphe disposant d'une unique composante connexe, sans toutefois s'intéresser à son diamètre (défini comme le plus long des plus courts chemins du graphe). Ce critère,

bien que primordial, ne nous semble pas suffisant. Nous proposons donc un ensemble de cinq propriétés intuitives auxquelles un graphe explorable devrait répondre.

Propriété 1 : Un arc entre deux nœuds indique que ces nœuds sont liés sémantiquement d'une façon ou d'une autre.

Propriété 2 : Il existe un chemin entre toute paire de nœuds.

Propriété 3 : Le plus court chemin entre deux nœuds arbitraires devrait être raisonnablement petit.

Propriété 4 : Il y a un nombre raisonnable de liens sortants pour chaque nœud.

Propriété 5 : La quantité de liens entrants est proportionnelle à l'importance du nœud.

La propriété 1 correspond à un critère de cohérence dans l'exploration. Bien que les liens créés puissent être discutables ou subjectifs comme explicité en section 4.1.1, il serait inopportun de créer un lien entre deux documents n'ayant rien à voir l'un avec l'autre, au risque de créer par là même une désorientation importante des utilisateurs (Sturgill et al., 2010).

La propriété 2 consiste à s'assurer que chaque document de la collection est accessible. Elle correspond au critère de *browsability* de Seyerlehner et al. (2009). Elle fait écho mathématiquement à l'existence d'une unique composante connexe dans le graphe. Plusieurs raisons expliquent ce choix. Tout d'abord, l'objectif consistant à permettre à l'utilisateur d'explorer de larges pans de la collection au travers du *berry-picking* nécessite une telle propriété. En effet, sans cela, un utilisateur pourrait rester bloqué au sein d'un petit nombre de documents discutant un sujet précis sans qu'aucune possibilité d'accès à une thématique connexe ne lui soit offerte, limitant de fait sa capacité d'exploration. De plus, on peut estimer que chaque document présent dans la collection étant issu du travail d'un ou plusieurs journalistes, il serait mal avisé de masquer son existence à l'utilisateur final. Cette considération se trouve renforcée par la constatation que la plupart des articles discutant un même événement apportent des informations uniques, qui ne se retrouvent nulle part ailleurs dans la collection (voir la section 8.2.2 pour plus de détails). Un corollaire de cette propriété est qu'il existe nécessairement au moins un lien entrant et un lien sortant pour chacun des nœuds du graphe.

La propriété 3 vise à éviter une exploration pénible de la collection dans laquelle, pour connecter deux documents arbitraires, un utilisateur aurait à parcourir plusieurs centaines de liens. D'un point de vue mathématique, il s'agit de s'assurer que le graphe ait un petit diamètre. Cette propriété disqualifie notamment l'organisation strictement chronologique d'une collection, qui consisterait en un long fil de documents, créant certes un graphe disposant d'une unique composante connexe, mais muni d'un diamètre correspondant au nombre de documents. Associée à la propriété 1 et 2, elle permet l'implémentation directe de fonctionnalités telles que celle proposée par Shahaf et Guestrin (2010), à savoir la possibilité d'offrir une suite de documents reliant deux informations choisies arbitrairement par l'utilisateur (par exemple, un ensemble de 5 documents démarrant par "Scandale de corruption à la FIFA" et terminant par "Sepp Blatter abandonne sa candidature à la tête de la FIFA").

La propriété 4 reflète l'attente des utilisateurs d'être guidés dans leur exploration. Même si un document donné pouvait être logiquement lié à plusieurs dizaines voire centaines d'autres documents, il est impraticable de demander à l'utilisateur de choisir quel prochain document il souhaite visiter parmi une si longue liste de possibilités.

Mathématiquement, cela revient à limiter le nombre de nœuds sortants, et donc le degré moyen, médian et maximum des nœuds du graphe. Une transposition directe sous forme de graphe des clusters proposés par les agrégateurs d'actualités violerait cette propriété étant donné que certains sujets sont très largement traités par une multitude de médias et regroupés au sein du même ensemble.

Enfin, la propriété 5 vise à s'assurer que les informations les plus discutées dans la collection soient effectivement aisément accessibles, c'est-à-dire qu'un large nombre de liens y mènent.

4.2.2 Différences avec la notion de navigabilité

Le terme « navigabilité » a été utilisé afin de décrire des graphes que nous appelons dans cette thèse « explorables », par exemple dans le domaine de la recommandation (Lamprecht et al., 2016). Or, la navigabilité fait également référence à une propriété mathématique des graphes, issue de l'analyse du « phénomène du petit monde » (Milgram, 1967). Ce phénomène a été d'abord observé par le biais d'études sociales indiquant que deux personnes prises au hasard dans la population sont reliées par une chaîne de relations (*i.e.* de connaissances) relativement courte, généralement estimée de 3 à 6 personnes. L'expérimentation consiste à fournir une lettre à une personne, avec la description d'un individu. L'objectif est alors, en passant cette lettre de connaissance en connaissance, d'arriver jusqu'à cet individu. Si ce phénomène indique bien qu'un graphe des relations entre individus dans une société serait composé d'une unique composante connexe au diamètre relativement court, et serait donc compatible avec notre terme d'explorabilité, la notion de navigabilité proposée par Kleinberg (2000) dans son analyse du petit monde va plus loin. En effet, il remarque que l'une des principales difficultés rencontrées dans les expérimentations de la théorie du petit monde consiste à identifier, pour chaque acteur, lesquelles de ses relations sont les plus à même de le rapprocher de l'individu objectif. Selon Kleinberg (2000), l'une des conditions nécessaires à ce phénomène est la présence de quelques liens « lointains » connectant deux personnes partageant des connaissances très différentes. Autrement dit, chaque individu évolue dans des cercles où les gens se connaissent les uns les autres, mais dispose également de quelques connaissances originales, qui ne sont pas partagées par ses cercles. Ce sont ces connaissances qui permettent d'atteindre plus facilement, en moins d'étapes, l'individu objectif.

Bien qu'il ait été prouvé que de nombreuses collections issues du web sont navigables (Adamic, 1999; Adamic et Adar, 2005), cette thèse ne défend pas l'idée que les graphes explorables doivent nécessairement répondre aux critères de navigabilité.

Conclusion

La construction d'hypergraphes d'actualités peut être envisagée d'au moins deux façons : comme une alternative à la recommandation, ou comme une structure permettant l'exploration d'une collection. Dans le premier cas, la structure de graphe n'est finalement que secondaire, et des évaluations sur la pertinence des liens créés favoriseront la suggestion de documents très populaires, quitte à encourager la dissimulation de la majorité des documents présents dans la collection au bénéfice de quelques-uns, jugés les plus pertinents. Dans le second cas, qui est celui défendu dans cette thèse, la structure d'hypergraphe peut être envisagée comme un moyen de découvrir la richesse d'une collection, en rendant chacun de ses éléments facilement accessible, sans pour autant rogner

sur la pertinence des liens créés. Pour cela, la notion d'explorabilité du graphe proposée ici semble essentielle.

Dans ce chapitre, nous avons défini précisément la notion d'hypergraphe explorable, ainsi que ce qu'elle impliquait. Nous pensons que dans une approche d'exploration d'une collection, sans besoin d'information précis, les caractéristiques souhaitables énoncées plus haut permettent un parcours efficace aux utilisateurs, sans risquer de les perdre au milieu de centaines de suggestions, et sans toutefois leur cacher aucun document. Les propriétés énoncées peuvent sembler contradictoires, certaines nécessitant un nombre de liens importants (unique composante, diamètre court) et d'autre les restreignant (contrainte de cohérence, degré faible) mais c'est leur équilibre qui doit permettre une navigation aisée au sein de grands corpus. L'intérêt de telles structures est démontré par des tests utilisateurs présentés dans le chapitre 8.

Chapitre 5

Construction de graphes explorables

Introduction

Malgré leur multiplicité, l'explorabilité respective des moyens disponibles pour consulter l'actualité n'a été que peu étudiée. Ceci s'explique en partie par une première difficulté : celle consistant à évaluer la pertinence d'un lien entre deux documents. En effet, sans cette contrainte de cohérence ou d'intérêt, il est aisé de construire une structure facilement explorable, au dépend de liens non pertinents. Or, cette évaluation est particulièrement complexe car, comme décrit précédemment, plusieurs types de relations entre deux éléments peuvent être envisagées, allant de critères stricts tels que la description d'un même événement, à des critères plus larges tels que le partage d'une thématique commune comme l'économie ou le sport, voire à des critères subjectifs tels que la causalité. Si les critères d'explorabilité décrits dans la section 4.2 peuvent être quantifiés, l'absence de corpus objectif permettant d'évaluer la pertinence des liens créés rend difficile la comparaison entre différentes méthodes de structuration d'actualités.

Au sein de ce chapitre, nous proposons une méthode permettant de résoudre pour partie cette problématique, et nous attachons à évaluer l'explorabilité de deux algorithmes standard de construction de graphes dans le contexte de collections d'actualités. Constatant leurs limites importantes en termes d'explorabilité, nous proposons une nouvelle approche permettant la construction de graphes d'actualités plus explorables, sans pour autant rogner sur la pertinence des liens créés.

Ce chapitre est organisé comme suit. Dans un premier temps, nous décrivons la méthodologie employée pour faire face à l'absence de corpus idéal, via l'utilisation d'une extraction de *clusters* de Google News fondés sur un regroupement des articles discutant d'un même événement. Nous continuons par l'étude des caractéristiques de deux méthodes classiques de construction de graphes fondées sur la notion de similarité entre les paires de documents : les K plus proches voisins (K-NN) et les epsilon plus proches voisins ($\mathcal{E}$-NN). Les deux méthodes construisent un graphe non orienté, dans lequel les nœuds correspondent aux documents, et les arcs à un lien entre une paire de documents. L'approche K-NN consiste, pour chaque document, à le lier aux K documents qui lui sont le plus similaire, K étant une constante définie pour l'ensemble du corpus, le plus souvent de façon arbitraire. La méthode $\mathcal{E}$-NN suit le même processus, mais utilise comme constante un seuil de similarité. Pour chaque paire de documents, si leur score de simi-

larité est supérieur à une constante $\mathcal{E}$, un lien est créé entre ces deux documents. Enfin, nous proposons une nouvelle approche, les plus proches voisins adaptatifs (*adaptive nearest neighbours*, ou A-NN), qui évite l'utilisation de seuils fixés manuellement, et dispose de caractéristiques d'explorabilité plus avantageuses que les approches K-NN et $\mathcal{E}$-NN. Nous discutons en détail les propriétés de cette nouvelle approche, en abordant notamment les thématiques de complexité algorithmique et les possibilités de mises à jour du modèle.

5.1 Cadre expérimental : des clusters à l'hypergraphe

L'évaluation de l'hyperliage est un problème complexe, et la proposition d'ajouter un critère d'explorabilité à l'hypergraphe résultant du processus complique encore davantage la tâche. Nous discutons dans cette section de la façon pertinente d'évaluer les critères d'explorabilité établis plus tôt. La notion de pertinence d'un lien, particulièrement problématique, y est largement abordée. Nous décrivons également le corpus sélectionné pour évaluer l'explorabilité des hypergraphes construits plus tard.

5.1.1 Protocole d'évaluation

Il n'existe pas, à notre connaissance, de corpus où chaque paire de documents aurait été annotée afin d'indiquer si les deux éléments sont liés ou non. Ceci s'explique par le fait qu'il est extrêmement difficile de construire un tel corpus. En effet, la pertinence du lien qui unit potentiellement deux documents dépend de l'information recherchée par l'utilisateur (Ge et al., 2010), et est le plus souvent subjective (Vargas et Castells, 2011). Si les liens unissant des documents discutant d'un même événement sont généralement perçus comme pertinents par les utilisateurs, l'intérêt d'établir un lien entre deux événements différents varie davantage (Bogers et Van den Bosch, 2007). Ainsi, un journaliste traitant de la corruption trouvera un lien entre l'affaire de la FIFA et d'autres affaires de corruption dans le milieu politique cohérent. À l'inverse, un journaliste cherchant à retracer l'historique de l'association FIFA ne sera pas intéressé par ce lien. Une seconde difficulté limitant la possibilité de créer un tel corpus est due à l'explosion du nombre de liens possibles pour chaque nouveau document ajouté à une collection. Ainsi, dans une collection composée de N documents, l'ajout d'un document supplémentaire nécessite l'annotation de N nouveaux liens potentiels.

Face à cette absence de corpus adapté et à l'extrême difficulté de sa construction éventuelle, nous avons choisi de développer une nouvelle méthodologie d'évaluation consistant à utiliser une vérité terrain fondée sur des *clusters* d'articles discutant un même événement. Ces *clusters* correspondent aux articles réunis au sein d'un même groupe par les agrégateurs d'actualité. Il est important de noter que chercher à approcher une vérité terrain de type *clusters* correspondrait à créer un graphe aux multiples composantes, dans lequel les articles d'un même groupe sont tous liés les uns aux autres, et liés à aucun document d'un autre *cluster*. Ceci ne correspond pas aux critères d'explorabilités décrits précédemment. Nous choisissons néanmoins d'utiliser un corpus de ce type afin de s'assurer de la proportion de liens objectivement corrects, qui correspondent aux liens créés entre deux documents discutant d'un même événement, et appartenant donc au même *cluster*. En effet, le choix d'apparier deux articles qui décrivent le même événement porte moins de subjectivité que les liens plus faibles que nous cherchons également à construire dans nos graphes. Ce corpus de *clusters* nous permettra donc d'évaluer la proportion de

liens créés qui correspondent à un lien intra-cluster, dont la justification ne fait pas de doute, et ceux qui correspondent à des liens extra-clusters, dont il sera impossible, sans annotation supplémentaire subjective, de savoir s'ils sont justifiés, mais qui permettront de répondre aux critères d'explorabilité souhaités pour notre graphe.

5.1.2 Caractéristiques du corpus

Le corpus utilisé est extrait d'une aspiration de quatre des catégories proposées par la version anglaise de l'agrégateur d'actualité de Google News [1] sur 5 mois pendant l'année 2015, réalisée par Gasparetti (2016) et disponible en ligne sur le site du UCI [2]. Il est composé d'une liste d'URL correspondant à des articles de presse en ligne, ainsi qu'à l'identifiant du cluster auquel ces articles ont été rattachés sur Google News. La table 5.1 présente quelques caractéristiques du corpus tel qu'il est distribué.

Catégorie	URL	Clusters	Taille moyenne des clusters
Santé	45 554	1 314	34,6
Sciences et Tech	108 324	1 779	60,9
Économie	115 889	1 994	58,1
Divertissement	152 461	2 070	73,7

TABLE 5.1 – Caractéristiques du corpus UCI.

De nombreuses URL fournies au sein de ce corpus ne sont plus accessibles. Ceci peut être causé, par exemple, par la limitation dans le temps de la disponibilité des articles pratiquée par certains journaux en ligne, ou bien par le changement des URL qui référencent ces articles. Nous avons récupéré l'ensemble des pages web qui pouvaient l'être, en s'aidant du site *The Internet Archive* [3], afin de recueillir les versions les plus anciennes de ces articles, tels qu'ils ont été publiés initialement. Nous avons ainsi pu obtenir environ 15 000 articles par catégorie, à savoir 15 689 pour la catégorie business, 14 509 pour la catégorie divertissement, 14 086 pour la catégorie santé, et 16 284 pour la catégorie sciences et technologies. Ces articles étant sous format HTML, avec menus de navigation et encarts publicitaires, il a fallu en extraire le contenu principal, c'est-à-dire le corps de l'article. Pour cela, nous avons utilisé la librairie Python Newspaper [4], qui se fonde sur des heuristiques afin de détecter les titres, auteurs, dates et textes présents sur les pages d'actualité en langue anglaise.

Il est intéressant de regarder les propriétés des *clusters* obtenus dans le cadre de la création de ce corpus, notamment leur homogénéité en termes de taille et de similarité. En d'autres termes, les groupes comportent-ils tous environ le même nombre de documents, et les documents appartenant à un même groupe sont-ils similaires ? La similarité peut se décrire de nombreuses façons (sémantique, lexicale, thématique, ...). Dans notre cas, nous nous intéressons à la similarité informationnelle et proposons de l'estimer à l'aide d'une représentation et d'une mesure très largement utilisées : le tf-idf et le cosinus, décrits plus tôt dans ce manuscrit. En effet, cette mesure de similarité lexicographique est en corrélation avec la quantité d'informations partagées par deux documents, au sens

1. http ://news.google.com
2. http ://archive.ics.uci.edu/ml/
3. www.archive.org/
4. www.github.com/codelucas/newspaper

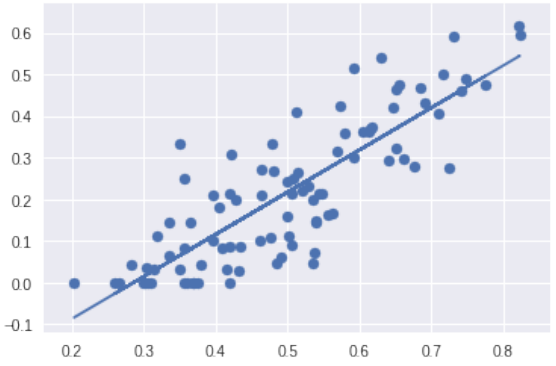


FIGURE 5.1 – Corrélation entre la part d'informations communes (abscisse) et la similarité lexicale (ordonnées).

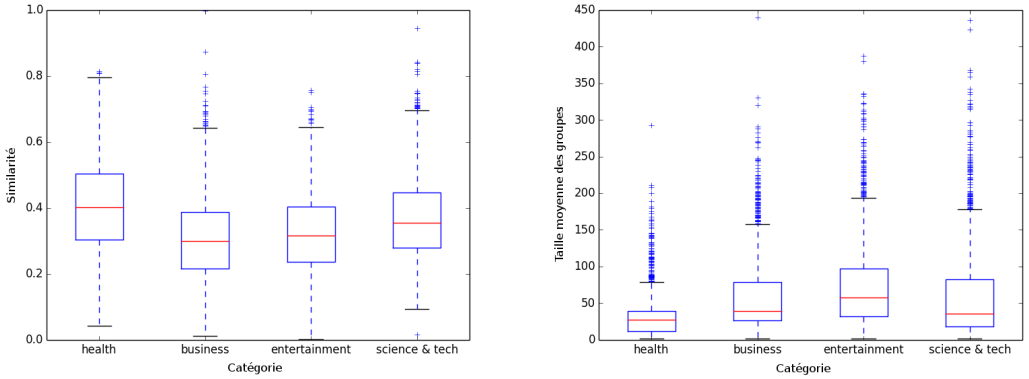


FIGURE 5.2 – Variabilité des similarités intra-clusters et des tailles de clusters selon les catégories ; à gauche la distribution des scores de similarité intra-clusters, à droite la distribution des tailles des clusters.

de l'information telle que définie plus tôt. Lors d'expérimentations mises en places afin de conduire des tests utilisateurs, nous avons extrait du corpus LIMAH l'ensemble des informations traitant l'avion solaire "Solar Impulse". La figure 5.1 montre que les scores obtenus via une représentation tf-idf associée à la mesure cosinus sont en corrélation avec le nombre d'informations partagées, calculé à l'aide d'une mesure de Jaccard (nombre d'informations partagées divisé par le nombre d'informations total).

La figure 5.2 montre les variations de taille et de similarité au sein des groupes, la similarité intra-groupe étant définie comme la moyenne des similarités entre paires de documents du groupe. On remarque que non seulement la taille des groupes varie énormément, avec une moyenne plutôt faible de l'ordre de quelques dizaines d'articles, mais pouvant aller de seulement quelques documents à plusieurs centaines. On note également que ces tailles changent en fonction des catégories considérées, la catégorie santé variant moins que la catégorie divertissement. De la même façon, les similarités telles que nous les avons mesurées montrent de grandes variations, certains groupes étant extrêmement homogènes (avec des similarités entre les articles de 0,8), et d'autres très hétérogènes (0,05 de similarité moyenne). Une fois encore, les similarités moyennes constatées varient en fonction de la catégorie, les documents issus du domaine de la santé étant plus similaires les uns aux autres que ceux issus de l'économie. Ces caractéristiques, si elles

ne sont pas surprenantes, illustrent la difficulté de concevoir des algorithmes s'adaptant aux particularités de chaque thématique, et plus encore aux particularités de chaque information rapportée.

5.2 K-NN et $\mathcal{E}$-NN, un paramétrage complexe et une explorabilité limitée

La construction de graphes explorables peut être envisagée de différentes manières. Il est possible de s'appuyer sur la création préliminaire d'un graphe dense (Yang, 1993), dont les arrêtes peuvent être supprimées itérativement afin d'en améliorer l'explorabilité (Fellows et al., 2011). Une deuxième approche consiste à s'appuyer sur des techniques fondées sur l'optimisation pour trouver les liens les plus pertinents (Backstrom et Leskovec, 2011). Une dernière méthode, qui est celle exploitée dans ce manuscrit, consiste à construire des graphes de plus proches voisins, dans lesquels chaque document est lié aux documents qui lui sont le plus similaires.

La création de graphes de plus proches voisins, fondée sur l'utilisation de mesures de similarité, a été largement étudiée. Parmi les approches classiques, deux méthodes très utilisées coexistent : les K-NN et les $\mathcal{E}$-NN. Dans cette section, nous décrivons ces méthodes et montrons qu'appliquées à la tâche d'hyperliage, elles n'offrent qu'une explorabilité limitée. Leur combinaison, par intersection ou par union, y est également considérée.

5.2.1 K-NN

L'algorithme des K plus proches voisins (K-NN) est un des deux algorithmes les plus utilisés pour construire des graphes, le second étant $\mathcal{E}$-NN, discuté dans la section suivante. K-NN consiste à relier chacun des documents d'un corpus aux K documents qui lui sont le plus similaire, K étant un seuil fixé le plus souvent de façon manuelle. Mathématiquement, étant donné une mesure de distance d et un seuil K, la construction d'un graphe des plus proches voisins peut être définie récursivement ainsi :

$$
\begin{aligned}
K\text{-NN}^0 &= \emptyset \\
K\text{-NN}^k &= K\text{-NN}^{k-1} \cup \big\{ (v_i, v_j) \mid d(v_i, v_j) < d(v_i, v_k), \\
&\qquad \forall v_i, v_j, v_k \in V^3, \\
&\qquad v_i \neq v_j \neq v_k,\ (v_i, v_j) \notin K\text{-NN}^{k-1} \big\} \quad .
\end{aligned}
$$

Dans notre cas, la mesure de similarité utilisée est le cosinus, appliqué sur une représentation tf-idf des documents. Nous choisissons cette représentation et cette mesure, très largement utilisées en recherche d'information et en traitement automatique des langues, pour leur robustesse et leur facilité d'interprétation. On peut passer de la mesure de similarité entre deux vecteurs $sim(v_1, v_2)$ à la distance entre deux vecteurs $d(v_1, v_2)$ en prenant son opposé : $d(v_1, v_2) = 1 - sim(v_1, v_2)$. Ces deux notions de distance et de similarité sont utilisées de façon indifférenciée dans le reste de ce manuscrit. Le graphe K-NN que nous construisons, comme l'ensemble des autres graphes qui seront discutés ensuite, sont non orientés, c'est-à-dire que s'il existe un lien qui relie le document v_i au document v_j, alors

K	Comp	|Comp|	|Comp| %	Diamètre	Degré	Précision	Rappel
2	540	9 174	58,4 %	74	2	78,7 %	14,2 %
4	130	13 626	86,8 %	34	4	71,9 %	21,3 %
6	51	14 743	93,9 %	22	6	67,6 %	27,5 %
8	31	15 025	95,7 %	17	8	64,3 %	33,0 %
10	24	15 153	96,5 %	15	10	61,4 %	38,0 %

TABLE 5.2 – Nombre de composantes connexes, taille de la plus grande composante connexe, ratio de nœuds du graphe appartenant à la plus grande composante connexe, diamètre, degré, précision et rappel en fonction du paramètre K. La catégorie do corpus UCI utilisée est business.

un lien inverse allant de v_j à v_i est construit. Ceci permet, lors de la navigation au sein de ces graphes, de pouvoir facilement retourner en arrière après avoir suivi un lien.

L'approche K-NN pose d'emblée plusieurs problèmes en termes d'explorabilité et de cohérence des liens créés. En effet, l'approche, consistant à relier un document à ses voisins les plus proches, est très dépendante de son seuil K, qui est appliqué de façon non différenciée à l'ensemble des documents de la collection. Un K trop élevé créera, pour certains documents, des liens non pertinents avec des documents très peu similaires, très éloignés dans l'espace de représentation utilisé. À l'inverse, un K trop faible créera trop peu de liens, limitant les possibilités de navigation au sein du graphe. La description que nous avons faite du corpus en section 5.1.2 montre la grande variabilité des tailles des groupes d'articles, et il est légitime de penser qu'une approche de K-NN, dans laquelle chaque article est lié à un nombre fixe de voisins, créera de nombreux liens non pertinents.

La table 5.2 présente les caractéristiques des graphes K-NN en fonction de K, calculés sur la catégorie business du corpus UCI décrit en section 5.1. Les informations présentées sont le nombre de composantes dans le graphe, la taille de la plus grande des composantes, la proportion de nœuds qu'elle comporte, le diamètre du graphe (défini comme le plus long des plus courts chemins), son degré médian, la précision et le rappel des liens créés. La précision est définie comme le rapport entre les liens pertinents créés et l'ensemble des liens créés. La pertinence d'un lien correspond au fait que les documents liés appartiennent au même groupe, tel que défini dans le corpus UCI. Comme décrit plus tôt, obtenir une précision maximale reviendrait à ne créer que des groupes disjoints, sans connexion entre les différents *clusters*, limitant par là même l'explorabilité. La précision nous sert donc davantage d'indicateur de la proportion de liens dont la pertinence ne fait aucun doute par rapport à l'ensemble des liens créés. Le rappel est défini comme le rapport entre les liens corrects créés et l'ensemble des liens corrects qui auraient pu être créés, toujours en prenant comme vérité terrain les groupes du corpus UCI. Un rappel parfait nécessiterait de lier toutes les paires d'articles appartenant à un même groupe. Certains groupes comptant plus de 400 membres, chacun d'entre eux devraient donc être relié à tous les autres, offrant plus de 400 suggestions de poursuite de navigation aux utilisateurs qui parcourraient ce graphe, une situation qui ne répondrait pas aux critères d'explorabilité définis plus haut. Nous cherchons donc davantage un compromis, dans lequel la part de précision serait suffisamment élevée tout en permettant l'obtention de la plus grande composante possible, sans diminuer trop fortement le rappel.

Comme le montre la table 5.2, la précision tombe relativement rapidement à mesure que K augmente. Ceci peut facilement s'expliquer par la création de liens entre docu-

$\mathcal{E}$	Comp	|Comp|	|Comp| %	Diamètre	Degré	Précision	Rappel
0.59	868	4 653	29,7 %	34	9	69,9 %	40,1 %
0.63	640	7 539	48,8 %	37	13	66,6 %	46,5 %
0.70	313	12 251	78,0 %	27	18	58,9 %	56,2 %
0.78	130	14 500	92,4 %	19	29	47,0 %	66,9 %
0.79	116	14 746	93,4 %	17	32	45,1 %	68,2 %

TABLE 5.3 – Nombre de composantes connexes, taille de la plus grande composante connexe, ratio de nœuds du graphe appartenant à la plus grande composante connexe, diamètre, degré, précision et rappel en fonction du paramètre $\mathcal{E}$. La catégorie du corpus UCI utilisée est business.

ments peu similaires, K ne discriminant pas en fonction de la distance mais du rang des documents. À l'inverse, afin d'obtenir une composante connexe suffisamment grande, il est nécessaire d'utiliser un K grand. Le rappel est relativement faible étant donné les petites valeurs de K utilisées. La valeur idéale de K semble être entre 4 et 6, une valeur supérieure baissant trop fortement la précision. Ces valeurs n'offrent néanmoins qu'un rappel relativement faible.

5.2.2 $\mathcal{E}$-NN

L'approche des $\mathcal{E}$ plus proches voisins (ou $\mathcal{E}$-NN) est la seconde approche la plus utilisée pour créer des graphes. De façon similaire à K-NN, elle consiste à relier chaque document du corpus à ceux qui lui sont le plus similaire. Le seuil n'est cette fois pas un nombre de voisins, mais un seuil de distance $\mathcal{E}$. Chaque nœud a donc un nombre de voisins spécifique, déterminé par le nombre de documents qui lui sont fortement similaires. Tout comme les K-NN, l'une des majeures difficultés de cette approche consiste à fixer le seuil $\mathcal{E}$. Mathématiquement, étant donné une mesure de distance d et une distance limite $\mathcal{E}$, les $\mathcal{E}$-NN peuvent être définis ainsi :

$$\mathcal{E}\text{-NN} = \{(v_i, v_j) \mid d(v_i, v_j) < \mathcal{E},\ \forall v_i, v_j \in V^2\ v_i \neq v_j\} \ .$$

De façon identique aux K-NN, nous utilisons la mesure cosinus et la représentation tf-idf afin d'obtenir des scores de similarité entre chaque paire de documents, et le graphe construit est non dirigé.

L'approche $\mathcal{E}$-NN est connue pour favoriser l'apparition de hubs (Weber et Monge, 2011; Radovanović et al., 2010), c'est-à-dire de nœuds hyperconnectés, reliés de façon disproportionnée à un grand nombre d'autres nœuds. De tels nœuds rendent évidemment l'exploration de la collection complexe, et il n'est pas raisonnable de proposer plusieurs centaines de liens vers des nœuds cibles à un utilisateur pour un seul nœud source. La table 5.3 présente les caractéristiques des graphes $\mathcal{E}$-NN en fonction de $\mathcal{E}$, calculés sur la catégorie business du corpus UCI. On remarque qu'afin d'obtenir une composante connexe suffisamment grande, il est nécessaire d'utiliser un seuil de distance élevé, supérieur à 0,70. Autrement dit, les documents liés seront peu similaires. Ceci se traduit par une précision qui chute rapidement, une composante contenant plus de 90 % des nœuds n'affichant qu'une précision inférieure à 0,50.

K et $\mathcal{E}$	Comp	|Comp|	|Comp| %	Diamètre	Degré	Précision	Rappel
$4 \cup 0.60$	108	13 833	88,1 %	28	10	64,5 %	44,2 %
$8 \cap 0.75$	249	12 260	78,1 %	29	8	67,5 %	32,1 %

TABLE 5.4 – Nombre de composantes connexes, taille de la plus grande composante connexe, ratio de nœuds du graphe appartenant à la plus grande composante connexe, diamètre, degré, précision et rappel en fonction des paramètres K et $\mathcal{E}$. La catégorie du corpus UCI utilisée est business.

5.2.3 Combinaisons de K-NN et $\mathcal{E}$-NN

Les approches K-NN et $\mathcal{E}$-NN peuvent être combinées en utilisant leur union ou leur intersection. Dans le cas de l'union, l'idée consiste à utiliser des valeurs de seuils relativement faibles, avec un nombre de voisins et une distance maximale faibles, créant de fait peu de liens. L'union des deux graphes, c'est-à-dire l'accumulation de l'ensemble des liens créés par chacune des deux méthodes, permet alors d'obtenir un nombre d'arcs plus important, tout en limitant le nombre de liens non pertinents créés lors de l'utilisation de seuils moins restrictifs. L'intersection, à l'inverse, consiste à ne conserver que les liens créés par les deux méthodes. On peut alors choisir des seuils moins restrictifs. Cette seconde approche semble particulièrement pertinente dans notre cas puisqu'elle permet de limiter dans une certaine mesure le degré des nœuds, tout en s'assurant de ne pas relier des documents qui seraient trop dissemblables.

La table 5.4 présente les caractéristiques des graphes obtenus sur la catégorie business du corpus UCI, avec des valeurs de K et de $\mathcal{E}$ optimisées pour obtenir un le meilleur compromis possible en termes de taille de composante principale et de précision. On remarque que les composantes principales sont de taille inférieure aux graphes K-NN et $\mathcal{E}$-NN pris individuellement, mais permettent globalement une meilleure précision des liens créés.

5.3 ANN, une méthode non paramétrique pour la construction de graphes explorables

D'après les expérimentations décrites plus tôt, le manque d'explorabilité des méthodes K-NN et $\mathcal{E}$-NN peut s'expliquer par l'utilisation d'un unique seuil K ou $\mathcal{E}$ pour l'ensemble de la collection. Nous proposons donc de nous affranchir de l'utilisation d'un tel seuil, et d'adapter la valeur limite de similarité engendrant la création d'un lien à chaque document de la collection. Nous nous fondons pour cela sur l'exploitation d'une caractéristique de l'espace de représentation employé, que nous décrivons en début de section. Nous poursuivons par la description de notre méthode permettant la création de graphes d'actualités avant de la comparer aux deux algorithmes standards $\mathcal{E}$-NN et K-NN selon les critères d'explorabilité. Nous continuons par une analyse manuelle de la pertinence des liens créés avant de proposer des optimisations du modèle permettant d'envisager une utilisation « en ligne » de l'approche dans laquelle des documents peuvent être continuellement ajoutés. Enfin, nous montrons que les approches fondées sur une représentation sémantique des documents à l'aide de réseaux de neurones ne permettent pas d'appliquer notre méthode.

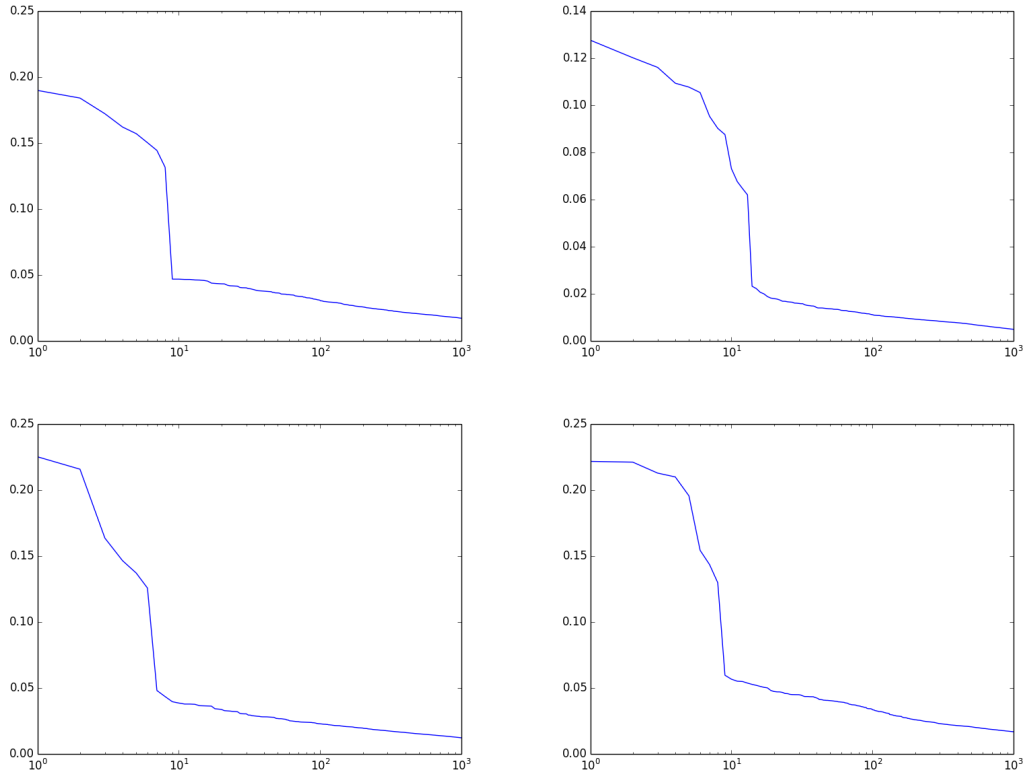


FIGURE 5.3 – Exemples de chutes de similarité sur la catégorie santé du corpus UCI. L'axe des abscisses indique le rang (échelle logarithmique), et l'axe des ordonnées la similarité cosinus fondée sur une représentation tf-idf.

5.3.1 Une exploitation des caractéristiques de l'espace de représentation

Comme nous l'avons vu précédemment, l'une des principales difficultés des méthodes K-NN et $\mathcal{E}$-NN consiste à établir un seuil optimal pour la collection étudiée. L'existence même d'un seuil optimal peut être débattue étant donné les grandes variations en termes de similarités intra-clusters décrites en section 5.1.2. Afin de répondre à cette limite, nous proposons de nous appuyer sur une caractéristique des espaces de représentation des documents. Au sein de ces espaces multidimensionnels, il existe une distinction nette entre le voisinage proche et le voisinage lointain de chacun des documents. Cette distinction naturelle est notamment employée dans le regroupement par densité (*density-based clustering*) (Kriegel et al., 2011). Elle a été mise en avant récemment par les travaux de Danisch et al. (2013), qui montrent qu'au sein de très grands graphes, des communautés peuvent être automatiquement détectées par ce procédé.

Un bon moyen pour visualiser ces différences de voisinage consiste à représenter, pour chaque document, sa similarité avec chacun des autres documents du corpus, et à ordonner ces similarités par ordre décroissant. On peut alors remarquer une brusque chute de similarité, permettant de distinguer le voisinage proche du voisinage distant. Les figures 5.3 et 5.4 montrent chacune quatre exemples de telles chutes de similarité, sur la catégorie santé pour la première, et sur l'ensemble des catégories UCI (c'est-à-dire le corpus dans sa totalité soit environ 60 000 documents) pour la seconde. La représentation utilisée est le tf-idf et la mesure est le cosinus, de façon identique aux expérimentations

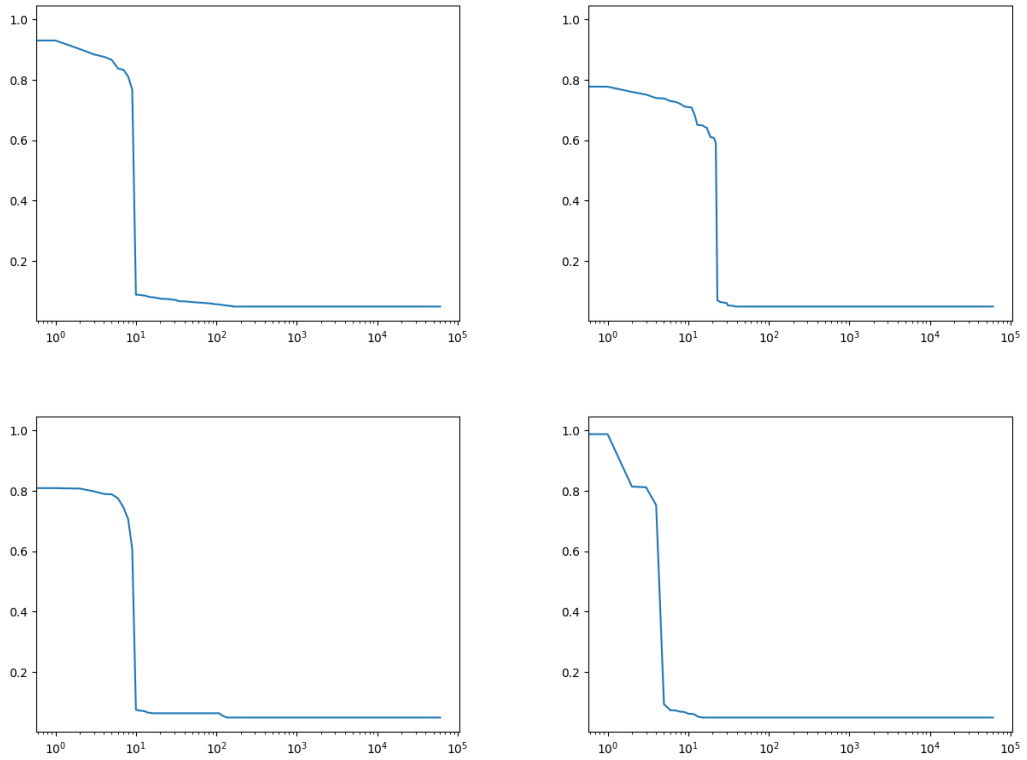


FIGURE 5.4 – Exemples de chutes de similarité sur l'ensemble des catégories du corpus UCI. L'axe des abscisses indique le rang (échelle logarithmique), et l'axe des ordonnées la similarité cosinus fondée sur une représentation tf-idf.

réalisées précédemment avec les K-NN et $\mathcal{E}$-NN. Il est important de noter que ces différentes chutes se produisent à des rangs différents (c'est-à-dire que le nombre de voisins proches diffère pour chaque document), et à des valeurs différentes (c'est-à-dire que le score de similarité auquel cette chute apparaît varie). Ceci répond à la principale critique faite à K-NN et $\mathcal{E}$-NN, à savoir la valeur fixée de leurs paramètres respectifs K (nombre de voisins) et $\mathcal{E}$ (seuil de similarité).

5.3.2 Méthode

L'approche que nous proposons, baptisée A-NN pour *Adaptive Nearest Neighbors* (plus proches voisins adaptatifs), consiste à utiliser les brusques chutes de similarité exposées précédemment afin de décider quels documents lier à un document d'intérêt. Ainsi, un document sera lié à son voisinage proche, apparaissant avant la chute de similarité. Ceci peut être interprété de deux façons distinctes, sémantiquement équivalentes. Une première façon, qui trace un parallèle avec les K-NN, consiste à dire que chaque document pourra donc avoir un nombre de voisins différencié en fonction de la quantité de documents apparaissant avant la chute. Une seconde façon, faisant davantage référence aux $\mathcal{E}$-NN, consiste à dire que chaque document sera lié aux documents dont le score de similarité sera supérieur au score où la chute de similarité apparaît. Mathématiquement, étant donné une mesure de la variation de similarité entre deux documents consécutifs Δ, et une mesure de distance d, les A-NN peuvent être définis ainsi :

Algorithm 1 *Adaptive nearest neighbors*

Input: V l'ensemble des documents ; v_i un élément de V
Output: E_i l'ensemble des voisins de v_i

```
l ← [ ]
for v_j ∈ V \ v_i do
    s ← sim(v_i, v_j)
    l.append(s)
end for
l ← sort(l)
k ← maxDropRank(l)                    (Voir Eq. 5.1)
E_i ← K-NN(V, v_i, k)
return E_i
```

$$\begin{aligned} \text{A-NN} &= \{(v_i, v_j) \mid d(v_i, v_j) < d_i,\ \forall v_i, v_j \in V^2\ v_i \neq v_j\} \\ d_i &= d(v_i, v_j) \mid v_j = argmax(\Delta_i(v_k, v_l)) \mid \forall v_k, v_l \in V^2 \\ &\qquad \text{t.q. } v_k, v_l \text{ sont des voisins successifs de } v_i \ . \end{aligned}$$

Cette définition est similaire à celle des $\mathcal{E}$-NN, à l'exception que le seuil d_i utilisé varie pour chaque document v_i et est déterminé par une détection de la plus forte baisse de similarité entre deux documents consécutifs Δ_i. Cette variation est exprimée simplement dans nos expérimentations par l'équation suivante :

$$\Delta_i(v_j, v_k) = 1 - \frac{\text{Sim}(v_i, v_k)}{\text{Sim}(v_i, v_j)} \ . \tag{5.1}$$

Cette formulation, plutôt qu'une simple différence de deux scores successifs, permet de diminuer l'importance d'une éventuelle chute arrivant très tôt, et le plus souvent dûe à des quasi duplicats (ou duplicats réels) dans les collections de ce type. En effet, grâce à cette formule, une chute de $0,4$ à $0,1$ est considérée comme plus importante ($\Delta_i = 0,75$) qu'une chute de $0,9$ à $0,5$ ($\Delta_i = 0,44$). Enfin, l'algorithme des A-NN peut être résumé, pour chaque document v_i en ces étapes successives :

1. Comparer v_i à chacun des autres documents de la collection.
2. Ordonner les documents comparés en fonction de leur score de similarité.
3. Détecter la plus forte chute de similarité d_i grâce à l'équation 5.1.
4. Créer un lien entre v_i et chaque document apparaissant avant d_i.

L'étape n° 1, quand appliquée à l'ensemble du corpus, consiste à calculer une matrice de similarité. Ce processus étant coûteux en termes de mémoire, les scores inférieurs à $0,05$ ne sont pas mémorisés. Ce choix a une importance à l'étape n° 3 puisque la plus forte chute de similarité n'est alors détectée que sur un sous-ensemble des voisins : ceux dont le score de similarité avec le document v_i est supérieur à $0,05$. Il n'existe dans le corpus qu'en moyenne cinq instances par catégorie (parmi les 15 000 instances de chaque catégorie) pour lesquelles aucune chute de similarité n'est détectée avant ce seuil (c'est-à-dire qu'il n'existe pas de document v_j tel que $sim(v_i, v_j) > 0,05$). Ces instances sont ignorées dans notre processus, c'est-à-dire qu'aucun lien n'est créé pour elles, tout comme

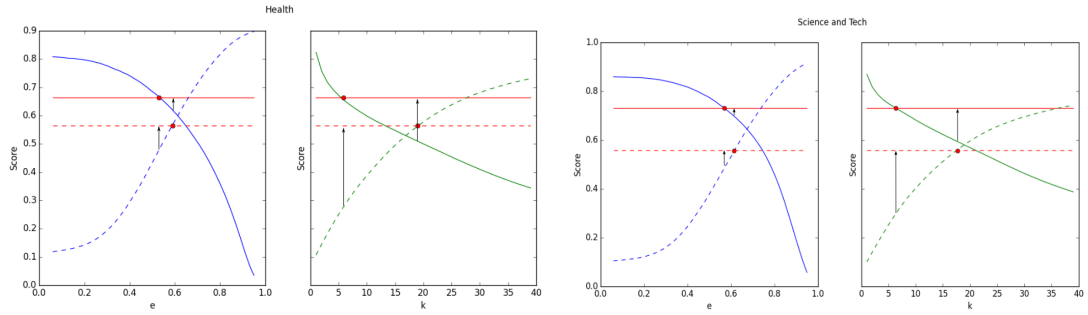


FIGURE 5.5 – Précision (lignes pleines), rappel (lignes pointillées) et gain en performance (flèches) pour A-NN (rouge), K-NN (vert), and $\mathcal{E}$-NN (bleu) sur les catégories santé (gauche) et science (droite) du corpus UCI.

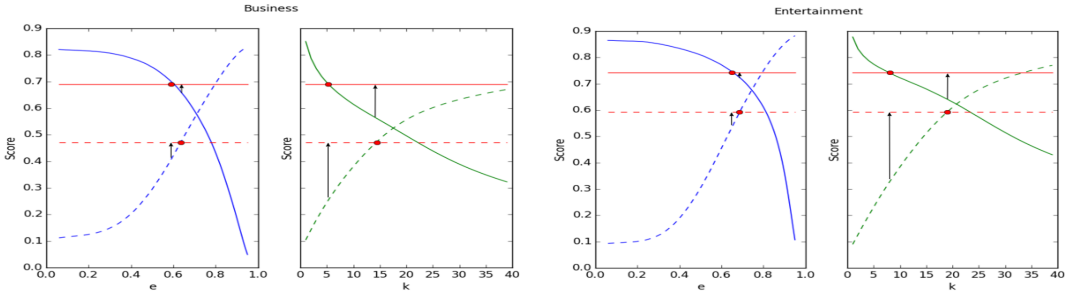


FIGURE 5.6 – Précision (lignes pleines), rappel (lignes pointillées) et gain en performance (flèches) pour A-NN (rouge), K-NN (vert), and $\mathcal{E}$-NN (bleu) sur les catégories business (gauche) et divertissement (droite) du corpus UCI.

c'est le cas pour l'approche $\mathcal{E}$-NN lorsqu'aucun document n'obtient de score de similarité supérieur au seuil $\mathcal{E}$. Ce choix de ne pas considérer les scores trop faibles nous prémuni également contre les valeurs trop faibles de $\text{sim}(v_i, v_j)$, qui font tendre Δ_i vers l'infini. L'approche des A-NN est récapitulé dans l'algorithme 1.

5.3.3 Comparaison de K-NN, $\mathcal{E}$-NN et A-NN

Nous pouvons comparer les trois méthodes étudiées selon deux angles : la différence de performance en termes de précision et de rappel des A-NN par rapport aux K-NN et $\mathcal{E}$-NN quels que soient leurs paramètres K et $\mathcal{E}$, et les différences topologiques des graphes créés. Les figures 5.5 et 5.6 montrent la première de ces comparaisons, à savoir le gain en précision (resp. rappel) pour un rappel (resp. précision) identique pour notre méthode non paramétrique et les approches de K-NN et $\mathcal{E}$-NN sur les quatre catégories du corpus UCI. Ainsi, sur la courbe la plus à gauche, on peut voir la précision de la méthode $\mathcal{E}$-NN chuter tandis que son rappel augmente (lignes bleues) en fonction de la valeur du seuil $\mathcal{E}$ (abscisse). Les scores de notre méthode sont illustrés par les lignes rouges et sont constants car ne dépendant pas de $\mathcal{E}$. La flèche noire de gauche indique le gain de notre méthode en termes de rappel pour une précision identique (point rouge de gauche). Inversement, la flèche de droite indique le grain de notre méthode en terme de précision pour un rappel identique (point rouge de droite). On remarque que sur chacune des quatre catégories, notre méthode améliore la précision (resp. le rappel), et que le gain est particulièrement grand par rapport aux graphes K-NN.

Catégorie	K / $\mathcal{E}$	Comp	|Comp|	|Comp| %	Diam	Deg	Précision	Rappel
Science	_	102	15 244	93,6 %	20	11	73,2 %	55,8 %
Business	_	110	14 954	95,3 %	17	8	69,7 %	45,8 %
Santé	_	74	13 314	94,6 %	20	13	66,4 %	56,5 %
Divert.	_	51	14 012	96,5 %	18	17	74,3 %	59,2 %
Science	6	126	14 401	88,4 %	27	6	73,5 %	28,7 %
Business	5	80	14 230	94,8 %	26	5	69,6 %	24,5 %
Santé	5	219	11 683	83,4 %	27	5	67,1 %	25,0 %
Divert.	8	33	13 915	95,9 %	19	8	74,2 %	32,6 %
Science	0,57	738	2 608	16,0 %	29	11	73,0 %	48,6 %
Business	0,60	799	5 512	36,7 %	16	10	69,2 %	42,2 %
Santé	0,53	681	680	4,8 %	13	15	66,6 %	47,8 %
Divert.	0,66	470	5 075	33,1 %	17	16	74,0 %	55,0 %

TABLE 5.5 – Nombre de composantes connexes, taille de la plus grande composante connexe, ratio de nœuds du graphe appartenant à la plus grande composante connexe, diamètre, degré, précision et rappel pour les graphes A-NN, K-NN et $\mathcal{E}$-NN.

La table 5.5 montre la seconde comparaison, et présente les caractéristiques des graphes créés par notre méthode sur chacune des quatre catégories du corpus UCI, ainsi que les caractéristiques des graphes K-NN et $\mathcal{E}$-NN, dont les paramètres sont optimisés pour offrir une précision similaire à notre approche. On remarque que notre approche offre systématiquement un rappel très largement supérieur aux deux autres approches, ainsi qu'une composante principale toujours plus grande. On remarque également que les valeurs optimales des seuils des graphes K-NN et $\mathcal{E}$-NN varient beaucoup selon les catégories, allant de $K = 5$ à $K = 8$ pour K-NN et de $\mathcal{E} = 0,53$ à $\mathcal{E} = 0,66$ pour $\mathcal{E}$-NN. Une fusion des catégories au sein d'une seule collection aboutirait donc à de nouveaux seuils optimaux pour K-NN et $\mathcal{E}$-NN, issus de compromis entre les différentes catégories. Ce n'est pas le cas pour A-NN, dont le seuil est défini indépendamment pour chaque nœud.

5.3.4 Validation sur le corpus LIMAH

Le corpus LIMAH est différent du corpus UCI sur plusieurs aspects. Tout d'abord, et principalement, le corpus LIMAH est multimodal, et dispose aussi bien d'articles de presse que d'émissions radiophoniques et télévisuelles. Ensuite, il est en langue française et non anglaise. Enfin, il est plus petit mais davantage contrôlé, la qualité des documents du corpus LIMAH ayant été vérifiée, notamment la bonne extraction du contenu principal des articles de presse écrite.

Afin de s'assurer de la bonne performance de l'algorithme des A-NN, nous l'avons appliqué au corpus LIMAH puis avons procédé à une annotation manuelle partielle des liens créés. Seule la pertinence des liens (*i.e.*, la précision de la méthode) a été évaluée, et non l'absence de documents potentiellement pertinents (*i.e.*, le rappel). Nous avons appliqué l'algorithme des A-NN sans aucune variation, bien que le contexte multimodal puisse justifier l'utilisation d'approches plus poussées (Favre et al., 2004). La table 5.6 donne les scores de précision discriminés selon le nombre de liens proposés afin de différencier la précision moyenne au sein de petits groupes d'articles de celle de grands groupes. La table 5.7 précise les scores en fonction de la modalité du document cible. On remarque que les scores de précision sont globalement supérieurs à ceux obtenus

Nombre de liens	Précision	Ratio
3	0.82	159/195
4	0.54	28/52
5	0.72	93/130
6	0.73	79/108
7	0.94	79/84
8	0.5	8/16
9	0.90	73/81
11	1.0	11/11
12	0.94	34/36
13	0.90	35/39
17	1.0	17/17
19	1.0	19/19
27	1.0	27/27
28	0.89	25/28
39	1.0	39/39
moyenne	0.82	727/882

TABLE 5.6 – Précision et rapport entre liens corrects et liens incorrects en fonction du nombre de liens créés.

sur le corpus UCI. Ceci peut s'expliquer par la présence de documents malformés dans le corpus UCI, et par l'échantillonnage réalisé pour l'étude manuelle. Logiquement, les liens vers des documents issus d'articles écrits sont plus robustes que ceux créés vers un podcast radio ou télévisuel. Ceci s'explique notamment par le taux d'erreur relativement important des transcriptions automatiques.

Modalité	Précision	Ratio
Presse	0.83	671/804
Radio	0.71	48/68
Télévision	0.70	7/10

TABLE 5.7 – Précision et rapport entre lien corrects et liens incorrects selon la modalité du document cible.

5.3.5 Optimisations et mises à jour du modèle

L'approche que nous décrivons pour les A-NN nécessite un grand temps de calcul, principalement à cause de l'étape de comparaison de chaque paire de documents. La mesure de similarité utilisée étant symétrique, seule une demi matrice de similarité est nécessaire pour obtenir les scores de similarité de l'ensemble des paires de documents. Néanmoins, le calcul de cette demi matrice reste trop important pour une très grande collection (plusieurs centaines de milliers de documents ou davantage). D'autre part, l'ajout d'un nouveau document à la collection implique de recalculer une nouvelle ligne de cette matrice, une opération en $O(N)$ qui deviendrait trop coûteuse à terme sur des collections mises à jour en temps réel.

Une solution consiste à utiliser une approche de recherche de plus proches voisins approximée (Indyk et Motwani, 1998; Muja et Lowe, 2009), soit sous la forme d'un $\mathcal{E}$-NN

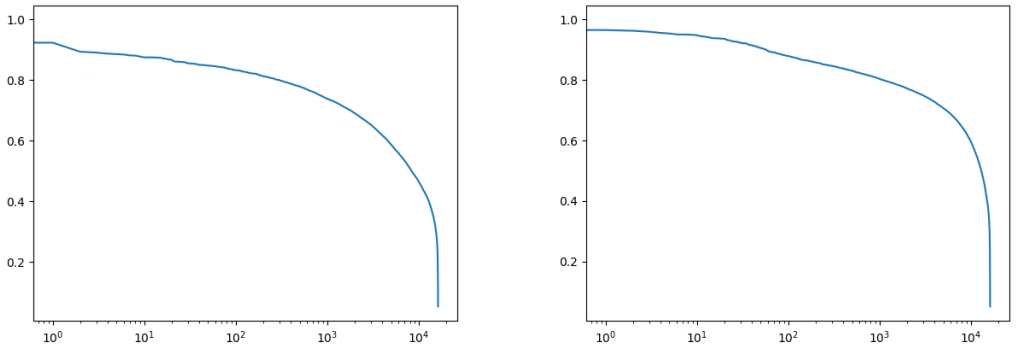


FIGURE 5.7 – Exemples de chutes de similarité sur la catégorie business du corpus UCI. L'axe des abscisses indique le rang (échelle logarithmique), et l'axe des ordonnées la similarité cosinus fondée sur une représentation Word2Vec.

approximé (il semble alors raisonnable de fixer un seuil de similarité à $0,05$, étant donné que la chute que nous cherchons à détecter apparaît presque toujours avant cette limite), soit sous la forme d'un K-NN approximé (rechercher quelques dizaines de plus proches voisins semble également justifié, la probabilité de manquer les vrais plus proches voisins devenant alors très faible). Une fois ces listes raccourcies de proches voisins obtenues, il suffit d'appliquer l'algorithme des A-NN afin de détecter la plus forte chute de similarité parmi ces voisins.

Une autre méthode pour obtenir cette liste raccourcie consiste à utiliser des approches de recherche d'information afin de trouver les documents les plus similaires (McCandless et al., 2010). De façon analogue aux K-NN approximés, il ne reste alors plus qu'à sélectionner quelques dizaines des résultats jugés les plus pertinents, et à appliquer l'algorithme des A-NN à ces documents.

Ces approches permettent d'envisager un calcul en direct (*on the fly*) du graphe, au fur et à mesure de sa consultation par les utilisateurs. Ainsi, à chaque déplacement sur un nœud, son voisinage approximatif peut être calculé et sa liste de voisins pertinents obtenue grâce à l'approche A-NN. Ceci permettrait alors d'éviter de stocker en mémoire le graphe complet. Avec ces méthodes approximées et un calcul du graphe en direct, la mise à jour du modèle devient triviale, étant donné que l'étape limitante de calcul d'une nouvelle ligne de la matrice de similarité n'est plus nécessaire. Il suffit alors pour chaque nouveau document d'obtenir sa représentation, et de l'ajouter en base.

5.3.6 Expérimentations sur la représentation neuronale de documents

Depuis quelques années, les représentations vectorielles profondes de mots et de documents, réalisés à l'aide de réseaux de neurones, apportent des améliorations de performances dans de nombreux domaines (Manning, 2016). Ces représentations vont plus loin que les représentations de surface, de type tf-idf, que nous avons utilisées plus tôt. En effet, elles permettent d'obtenir une représentation sémantique plutôt que lexicographique. Ainsi, dans l'espace de représentation construit, les synonymes, hyponymes ou métonymes se situent à proximité les uns des autres. À l'inverse, avec les représentations de surface, on ne saurait détecter une similarité sémantique que si le vocabulaire utilisé est identique. Les représentations Word2Vec et Doc2Vec sont deux exemples emblémа-

tiques de cette technique (Mikolov et al., 2013).

Nous avons souhaité appliquer l'algorithme des A-NN à de telles représentations, afin de bénéficier de représentations sémantiques plutôt que lexicographiques. Nous avons donc utilisé une représentation des documents en choisissant la moyenne des représentations Word2Vec des mots qui les composent. Pour cela, la librairie Python gensim [5] a été retenue. Un cosinus a ensuite été appliqué à ces représentations afin d'obtenir les scores de similarités pour chacune des paires de documents. Cette approche consistant à moyenner les vecteurs de mots s'est révélée efficace dans d'autres expérimentations de type ordonnancement (*ranking*) décrites en section 6.2.3. Malheureusement, comme le montre la figure 5.7, les chutes de similarité n'apparaissent pas lorsque de telles représentations sont utilisées, rendant l'algorithme des A-NN impraticable.

Conclusion

L'évaluation objective des performances de différents algorithmes pour la création de graphes explorables est un problème complexe. De nombreux critères sont à prendre en compte tels que l'accessibilité des différents éléments de la collection, la proportion de liens pertinents créés, ou leur nombre. Dans ce chapitre, nous avons proposé une nouvelle méthode pour évaluer les graphes créés sur de grandes collections, au travers de métriques mesurant la topologie du graphe, ainsi que l'utilisation de *clusters* d'agrégateurs afin d'évaluer la pertinence des liens créés. Nous avons également proposé une nouvelle méthode, les A-NN, pour construire des graphes explorables. Celle-ci se fonde sur une caractéristique des espaces de représentation utilisés, qui permet de différencier automatiquement le voisinage proche d'un document de son voisinage lointain, sans nécessiter l'utilisation d'un seuil.

Les métriques proposées permettent, dans notre cas, de montrer la supériorité des A-NN sur les deux algorithmes classiques de construction de graphes, les K-NN et $\mathcal{E}$-NN, les A-NN obtenant à la fois un meilleur ratio précision/rappel, et de meilleures performances en terme d'explorabilité. Il paraît néanmoins difficile d'imaginer que ces métriques permettent de différencier des algorithmes aux variations de performances moins flagrantes. Pour cela, seules des comparaisons extrinsèques à l'aide de tests utilisateurs, telles que celles réalisées chapitre 8, semblent adaptées.

5. www.github.com/RaRe-Technologies/gensim

Chapitre 6

Une diversité de liens nécessaire

Introduction

La pertinence des arcs créés, définie comme le fait que les liens soient sémantiquement cohérents, n'est pas le seul critère à observer lors du processus d'hyperliage. En effet, pour chacun des documents, ce sont plusieurs suggestions qui sont générées. Il convient alors d'étudier la diversité offerte par ces suggestions, créer plusieurs liens vers des documents extrêmement similaires d'un point de vue informationnel étant peu pertinent.

Dans ce chapitre, nous étudions l'intérêt et les moyens d'apporter une plus grande diversité dans les liens proposés lors d'une tâche d'hyperliage. Les différents systèmes qui y sont présentés ont été exploités dans le cadre des campagnes d'évaluation TRECVid et n'ont pas été mises en œuvre sur le corpus LIMAH. Ils constituent néanmoins des pistes pertinentes pour l'apport d'une plus grande diversité des liens dans le contexte de la construction d'hypergraphes navigables.

Nous débutons ce chapitre avec une discussion sur les avantages de l'obtention de liens divers et sur les différents types de diversité qui existent. Nous décrivons ensuite deux systèmes conçus dans le but d'améliorer la diversité des liens créés dans une tâche d'hyperliage, sans pour autant sacrifier à leur pertinence. Le premier système, introduit par Simon et al. (2015), se fonde sur une représentation des *topics* d'un segment vidéo selon les modalités visuelles et textuelles. Le deuxième système, introduit par Vukotić et al. (2016), est un réseau de neurones de type *autoencoder* permettant la construction d'un espace de représentation commun aux modalités visuelles et textuelles. Ces deux méthodes ont pour point commun de créer un espace de représentation commun à deux modalités, et de permettre la comparaison d'une modalité avec une autre. Elles sont donc qualifiées de *crossmodales*, une propriété pouvant notamment servir à améliorer la diversité des liens créés. Nous avons utilisé ces deux systèmes à l'occasion des campagnes de TRECVid 2015 et 2016 et avons réalisé une étude comparative de leur diversité par le biais d'une évaluation utilisateurs. Nous rapportons ici les performances obtenues lors des campagnes ainsi que les résultats de l'étude portant sur leur diversité. Enfin, nous concluons ce chapitre avec la description d'un troisième système, également fondé sur une représentation des *topics*, et dont l'objectif consiste à mieux maîtriser le degré de diversité souhaité lors du processus d'hyperliage. Notre contribution à ce dernier système

réside essentiellement dans la maîtrise de la topologie de la structure intermédiaire qu'il construit.

6.1 Les avantages de la diversité

La diversité offerte par des systèmes d'hyperliage, de recherche d'information ou de recommandation n'est pas un concept évident à définir, et elle prend dans ces trois tâches un sens différent. Dans cette section, nous décrivons les différents aspects de la diversité et les illustrons au travers d'exemples. Nous nous attachons particulièrement à deux aspects distincts offerts par la diversité : la capacité à répondre efficacement à des besoins de recherche imprécis, ambigus et implicites, et la sérendipité.

6.1.1 Des intérêts divers à concilier

Les approches décrites plus tôt dans cette thèse ont pour objectif la construction automatique de structures cohérentes au sein de collections d'actualité. Cette structure, unique à chaque collection, doit donc pouvoir répondre aux besoins de différents utilisateurs, quels que soient leurs besoins informationnels. Ainsi, deux utilisateurs consultant un même document peuvent être intéressés par des aspects tout à fait différents. Dans l'exemple du scandale de la FIFA, un premier utilisateur peut être intéressé par les enjeux politiques du scandale (conséquences sur les instances nationales, élection du président de la FIFA, ...) tandis qu'un second peut vouloir se focaliser sur les aspects légaux (condamnations encourues, conditions d'extradition, ...). Répondre de façon simultanée à ces deux besoins informationnels, ainsi qu'à la multitude d'autres possibles, est l'un des verrous scientifiques de la construction d'une structure fixe de navigation.

Le problème consistant à définir le besoin d'information d'un utilisateur a été largement exploré dans le cadre de la recherche d'information. Dans le contexte de l'utilisation de moteurs de recherche, l'utilisateur est le plus souvent invité à exprimer son besoin sous forme de mots-clés (Wang et al., 2006). Si des mots-clés initiaux peuvent servir à formuler ce besoin, il est souvent nécessaire à l'utilisateur d'itérer afin de trouver ceux qui seront les plus pertinents et les plus à même de répondre à son besoin. Dans ce contexte, deux des principales problématiques qui se posent sont la redondance (le moteur de recherche renvoie des résultats trop similaires les uns avec les autres) et l'ambiguïté (les mots-clés employés par l'utilisateur ne permettent pas d'expliciter clairement son besoin, *e.g.*, « couleur jaguar » peut faire référence à la couleur de l'animal, d'une voiture, voire du logo de la marque) (Clarke et al., 2008). Une façon d'y répondre consiste à fournir une liste de résultats diversifiés. Ainsi, afin de répondre efficacement à une requête ambiguë, plusieurs résultats correspondant chacun à une interprétation possible de la requête peuvent être proposés (*e.g.*, un document sur les voitures jaguar et un document sur l'animal). De la même façon, il est possible de limiter la redondance informationnelle des documents suggérés en s'assurant de retourner un ensemble de documents peu similaires, au moins sur le plan lexicographique.

Dans le cas de l'hyperliage, les enjeux rencontrés sont similaires, à l'exception du fait que le besoin de l'utilisateur n'est pas explicite. On peut en cela rapprocher l'hyperliage de la recommandation. La nécessité de la diversité dans les résultats des algorithmes de recommandation a été largement étudiée (Vargas et Castells, 2011), et cette problématique y est généralement solutionnée par l'exploitation du profil de l'utilisateur, qui tient peu ou prou le même rôle que les mots-clés des moteurs de recherche. Dans notre cas, nous

souhaitons construire des structures fixes, indépendantes du profil de l'utilisateur qui les explorerait. Étant donné l'impossibilité de s'appuyer sur une description du besoin informationnel ou sur des profils, il est essentiel de s'assurer la couverture la plus large possible des intérêts potentiels des utilisateurs.

6.1.2 La sérendipité

La sérendipité correspond à un résultat inattendu, un « hasard heureux » qui apporte une information pertinente bien que ne correspondant pas à l'attente originelle de l'utilisateur qui y est confronté (Bordino et al., 2013), et pouvant révéler des intérêts inconnus de l'utilisateur lui-même (Kamahara et al., 2005). La construction de systèmes apportant dans leurs résultats une dose de sérendipité a été explorée dans plusieurs domaines, notamment en recherche d'information et en recommandation (Foster et Ford, 2003).

Si ces résultats inattendus et pertinents sont souhaitables, ils sont néanmoins difficiles à évaluer, leur pertinence étant souvent subjective (Iaquinta et al., 2008). Ge et al. (2010) proposent d'évaluer la sérendipité de systèmes de recommandation en comparant les résultats qu'ils fournissent aux résultats d'un système primitif. Les résultats du système de recommandation n'apparaissant pas dans le système primitif sont alors considérés comme sérendipitueux, à la condition qu'un juge humain les déclare pertinents. Si ce moyen d'évaluation semble pertinent, de nombreuses problématiques d'ordre pratique apparaissent lors de sa mise en place, parmi lesquelles la définition du système primitif et l'évaluation humaine, subjective.

6.2 Fusionner les modalités pour une diversité plus large : LDA bimodal et réseau de neurones bimodal

Plusieurs moyens existent afin de limiter la redondance dans les résultats proposés par un moteur de recherche ou un outil d'hyperliage. On peut ainsi généralement adapter les algorithmes les plus couramment utilisés afin d'interdire la suggestion de documents trop similaires à un document déjà suggéré. Les expérimentations menées dans ce chapitre, ainsi que la plupart des travaux de cette thèse, se situent dans un contexte multimodal. Il est possible de tirer partie de cette caractéristique des corpus considérés, et de construire des systèmes visant à améliorer la diversité par le biais de la prise en compte simultanée des différentes modalités à notre disposition. Ceci peut se faire au travers de méthodes dites multimodales, comme par exemple la combinaison linéaire de scores de similarités fondées sur différentes modalités. Plus encore que la multimodalité, la crossmodalité, dans laquelle un espace de représentation commun à plusieurs modalités est construit, permet d'envisager une diversité accrue des systèmes qui la mettent en œuvre. C'est cette approche que nous défendons au sein de cette thèse. Dans cette section, après avoir discuté des avantages théoriques de la crossmodalité, nous présentons deux modèles crossmodaux conçus pour la tâche d'hyperliage et dont l'objectif est de trouver un compromis entre diversité et pertinence des liens créés.

6.2.1 Monomodalité, multimodalité et crossmodalité pour l'hyperliage

La tâche d'hyperliage (ou *hyperlinking*), décrite de façon générale en section 2.2.4, consiste à trouver au sein d'une grande collection multimédia, pour des segments

sources, une liste de segments cibles pertinents. Cette tâche fait l'objet d'une campagne d'évaluation annuelle à TRECVid, à laquelle nous avons régulièrement participé. Lors de ces évaluations, seule la pertinence des hyperliens créés, jugée par un unique vote via la plateforme Mechanical Turk, est prise en compte. Néanmoins, comme décrit plus tôt, l'hyperliage, par définition indépendant de l'utilisateur, nécessite une diversité importante dans ses résultats afin de s'adapter aux besoins variables des utilisateurs. Or, la plupart des systèmes proposés lors des campagnes d'évaluation TRECVid se fondent sur des similarités de surface (Guinaudeau et al., 2012a; De Nies et al., 2013; Bhatt et al., 2013; Le et al., 2014; Galuscáková et al., 2014; Barrios et al., 2015; Cheng et al., 2015; Pang et Ngo, 2015), utilisant une ou plusieurs modalités (image et/ou son principalement), et obtiennent généralement de bons scores de pertinence au détriment de la diversité des liens proposés. L'un des principaux axes d'amélioration du processus d'hyperliage réside donc en l'augmentation de la diversité devient.

Un moyen d'augmenter la diversité dans une tâche d'hyperliage multimédia consiste à considérer de façon simultanée plusieurs modalités. Un système monomodal aura en effet tendance à retrouver, pour un segment source, des segments cibles extrêmement similaires sur la modalité exploitée (*e.g.*, la parole ou l'image). Un système multimodal aura tendance à limiter la redondance des segments, en apportant une diversité plus importante du fait que la similarité puisse être dûe à l'une des modalités uniquement. Ainsi, dans un même système multimodal, certains segments proposés peuvent être similaires au segment source d'un point de vue langagier, et d'autres d'un point de vue visuel.

La crossmodalité, qui consiste à réunir plusieurs modalités au sein d'un même espace de représentation, peut potentiellement apporter encore davantage de diversité en permettant des similarités croisées. Ainsi, il est envisageable de construire des systèmes proposant des segments cibles dont l'une des modalités est similaire à une modalité différente dans le segment source. Ceci correspond par exemple à lier une vidéo montrant un universitaire parlant de châteaux à une vidéo montrant des châteaux.

Dans le cadre de la tâche *Hyperlinking* de TRECVid, les modalités exploitables correspondent essentiellement à la parole, au travers de transcriptions automatiques fournies par le LIMSI (Gauvain et al., 2002), et à l'image, grâce à des concepts visuels fournis par EUROCOM et fondés sur le modèle d'ImageNet (Krizhevsky et al., 2012). Lors de toutes les expérimentations qui suivent, les transcriptions automatiques ont été lemmatisées et les mots outils supprimés. Les collections vidéo utilisées dans les expérimentations qui vont suivre correspondent aux éditions 2015 et 2016 de TRECVid, soit respectivement à 3 mois de programmes de la BBC (environ 2 700 heures de vidéo) (Over et al., 2015) et à 14 838 vidéos de la plateforme BlipTV de 13 minutes en moyenne (Awad et al., 2016). Chaque programme est fourni comme une vidéo indépendante, allant de quelques minutes à quelques heures. L'hyperliage ayant pour objectif de suggérer des documents en lien direct avec un court passage d'une vidéo, une segmentation supplémentaire doit être appliquée afin d'éviter un lien vers un document trop long. De façon plus pragmatique, cette segmentation est une étape imposée dans le cadre de la tâche proposée à TRECVid, pour laquelle un système doit proposer des liens vers des segments d'au plus 120 secondes. Plusieurs segments correspondant à des fragments de vidéos de la collection et appelés ancres (100 pour le corpus 2015 et 90 pour le corpus 2016) sont fournis par les organisateurs. Pour chacune de ces ancres, les participants doivent trouver au sein de la collection une liste de segments pertinents, appelés cibles. La pertinence est évaluée, comme précisé plus tôt, par une annotation humaine réalisée via Amazon Mechanical Turk.

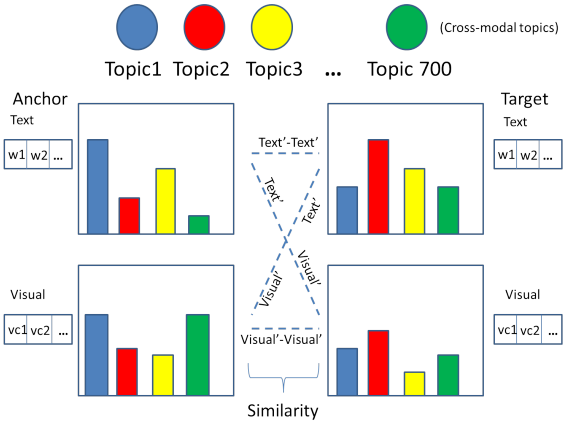


FIGURE 6.1 – Modèle CrossLDA.

6.2.2 LDA crossmodal

Les modèles *latent Dirichlet allocation* (ou LDA), permettent une représentation en thématiques (*topics*) latentes des documents. Cette méthode, très largement utilisée sur la modalité textuelle (Blei et al., 2003), offre une comparaison des documents au niveau des *topics* qu'ils abordent, plutôt qu'à un niveau lexicographique de surface. Les *topics* considérés correspondent chacun à une distribution de termes, appris sur un grand corpus. Ainsi, un *topic* pour lequel les termes les plus probables sont « casserole », « repas », « chef », « nourriture » et « frigo » correspond à une thématique liée à la cuisine. Les documents, plutôt que d'être représentés par les mots qui les composent, sont représentés par la distribution de leurs *topics*, notée θ . C'est donc cette distribution qui est utilisée, en conjonction avec des mesures de similarité, pour comparer les documents deux à deux. Cette approche a également été exploité sur l'image avec succès (Sivic et al., 2005). Le LDA bilingue (ou BiLDA, pour *Bilingual LDA*) a été introduit par Vulic et al. (2013, 2015) afin de solutionner des problématiques multilingues (*e.g.*, *clustering* multilingue, traduction). Leur approche propose de considérer des paires de documents comparables (qui ne sont pas en relation de traduction directe mais parlent de thématiques similaires) et d'apprendre les distributions de *topics* en forçant un parallélisme entre les distributions θ de chaque paire de documents comparables. Ainsi, après avoir appris le modèle, les représentations des deux documents parallèles (ou comparables) par leur distribution de *topics BiLDA* devraient être très similaires.

Nous exploitons cette idée de BiLDA en l'appliquant à deux modalités plutôt qu'à deux langues (Simon et al., 2015) [1] dans le cadre de l'hyperliage. Ainsi, pour un même segment vidéo, nous extrayons deux modalités : la modalité langagière, issue de la transcription automatique fournie par le système du LIMSI, et la modalité visuelle, issue des concepts visuels fournis par EUROCOM. Le modèle que nous utilisons n'est plus bilingue mais bimodal, et même crossmodal puisqu'il permet de passer d'une modalité à l'autre. Nous l'appelerons donc CrossLDA dans la suite de ce manuscript. Ainsi, après apprentissage des *topics*, pour un nouveau segment, nous pouvons extraire l'une des deux modalités (*e.g.*, le texte issu de la transcription), obtenir sa représentation CrossLDA, et chercher les documents ayant une distribution similaire, soit sur cette modalité textuelle, soit sur la

1. Ce système a été imaginé par Anca Simon, membre de l'équipe LinkMedia. Notre contribution a consisté à le mettre en œuvre dans le cadre de la campagne d'évaluation TRECVid 2015 et à évaluer sa diversité.

modalité visuelle. La figure 6.1 illustre les différentes combinaisons disponibles en fonction des modalités comparées.

Nous entraînons le CrossLDA à l'aide d'un échantillonnage de Gibbs avec les hyperparamètres $\alpha = 50/K$ et $\beta = 0.01$ (Steyvers et Griffiths, 2007), où K correspond au nombre de *topics* appris, dans notre cas 700. L'entraînement nous fournit une distribution de mots ϕ pour la modalité langagière et ψ pour la modalité visuelle. Ces distributions nous permettent d'estimer la contribution de chaque mot (resp. concept visuel) à un *topic* z_j. Étant donné une représentation textuelle (resp. visuelle) d'un document avec un vocabulaire V_1 (resp. V_2), la probabilité qu'un mot w_i (resp. concept visuel c_i) soit générée par le *topic* z_j est obtenue par la formule :

$$p(w_i|z_j) = \phi_{j,i} = \frac{n_{z_j}^{w_i} + \beta}{\sum_{x=1}^{|V_1|} n_{z_j}^{w_x} + \beta|V_1|} \tag{6.1}$$

où $n_{z_j}^{w_i}$ correspond au nombre de fois où le *topic* z_j a été assigné à une occurrence de w_i dans la phase d'apprentissage. La somme du dénominateur correspond au nombre d'occurrences total des mots assignés à z_j et β est l'a priori de Dirichlet. La formule 6.1 s'applique de la même façon aux concepts visuels avec $p(c_i|z_j) = \psi_{j,i}$. Ce modèle permet alors d'obtenir une représentation dans l'une ou l'autre des deux modalités pour n'importe quel document en calculant la probabilité du segment vidéo étant donné chacun des *topics* appris. La distribution obtenue est ensuite normalisée pour sommer à 1 afin d'obtenir la probabilité de chaque *topic* pour un segment. Cette normalisation est réalisée pour un document d selon la formule :

$$p(d|z_j) = (\prod_{i=1}^{n_d} p(w_i, z_j))^{\frac{1}{n_d}} \tag{6.2}$$

où n_d correspond à la taille du vocabulaire de d. Dans la pratique, nous calculons pour chaque document d deux représentations, la première correspondant à la modalité visuelle (notée d^v) et la seconde à la modalité langagière (notée d^l). Ces deux représentations peuvent être comparées avec celles d'un second document indépendamment des modalités considérées. En d'autres termes, pour deux documents d_1, d_2 et leurs représentations respectives, il est possible de comparer d_1^l à d_2^l, mais également d_1^l à d_2^v, tel qu'illustré en figure 6.1. La mesure de similarité que nous avons choisie pour la comparaison des distributions de deux documents est la mesure cosinus.

La table 6.1 montre le parallélisme des *topics* appris sur le corpus de TRECVid 2015. On remarque, pour le *topic* 7, que des mots tels que *food, cook, kitchen* sont logiquement associés à des concepts visuels liés à des aliments (*fig, pumpkin, zucchini*). Pour le *topic* 3, les concepts visuels probables correspondent à de la musique (*concert, singer, microphone*). La musique étant souvent accompagnée de paroles chantées, on retrouve des termes apparaissant fréquemment lors de concerts (*love, feel, like, baby*).

6.2.3 Réseaux de neurones bidirectionnels

Bien que les modèles fondés sur LDA soient très répandus, les améliorations récentes apportées dans de multiples domaines par l'utilisation de réseaux de neurones profonds encouragent l'adoption de cette nouvelle technique. Les réseaux de neurones bidirectionnels (ou BiDNN pour *BiDirectionnal Neural Networks*), introduits par Vukotić et al.

Topic 3	mots	love, home, feel, life, baby
	concepts visuels	singer, microphone, sax, concert, flute
Topic 7	mots	food, bit, chef, cook, kitchen
	concepts visuels	fig, acorn, pumpkin, guava, zucchini
Topic 25	mots	years, technology, computer, key, future
	concepts visuels	tape-player, computer, equipment, machine, appliance

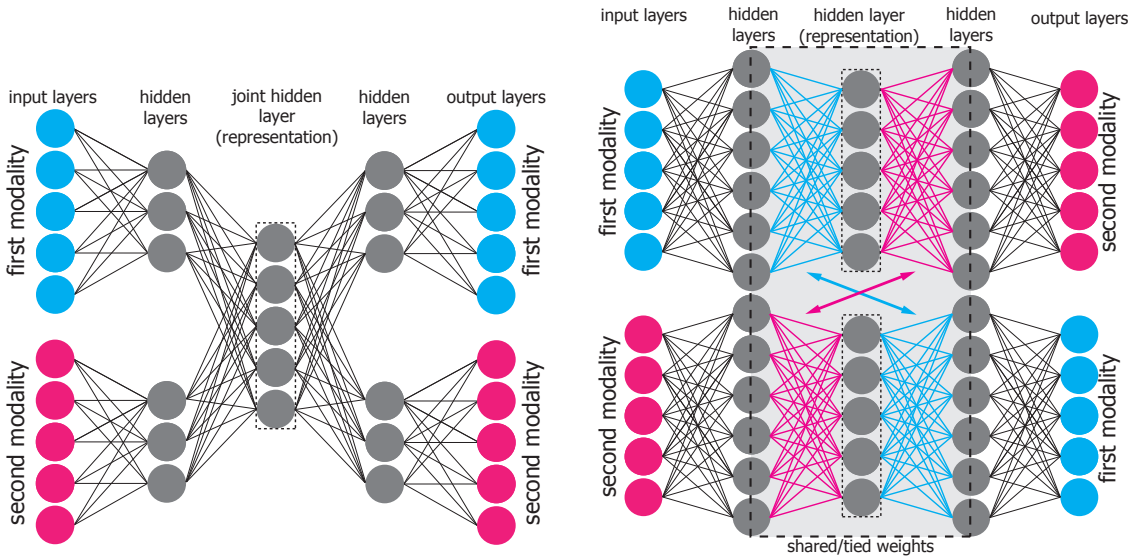
TABLE 6.1 – Les mots et concepts visuels les plus probables dans trois *topics* appris par le modèle CrossLDA.

(2016) [2], sont similaires à l'idée exploitée par le BiLDA, et correspondent à une architecture particulière d'*autoencoders*. Les *autoencoders* sont des réseaux neuronaux permettant de compresser l'information et fournissent des représentations pertinentes pour de nombreuses tâches, dont la recherche d'information multimédia (Hinton et Salakhutdinov, 2006). Il s'agit de réseaux prenant en entrée une représentation d'un document et ayant comme objectif de fournir en sortie une représentation identique. L'intérêt vient de la présence d'une couche neuronale centrale de taille inférieure aux couches d'entrée et de sortie. C'est cette couche centrale, plus petite, qui sert à encoder l'information et qui sera utilisée comme représentation pour chaque document. Les BiDNN permettent, selon le même principe de compression de l'information, de combiner deux modalités au sein d'un même espace de représentation. Plusieurs architectures prennent en compte les aspects multimodaux de certaines collections, et proposent une représentation crossmodale fondée sur des *autoencoders* (Feng et al., 2014). Une architecture classique est présentée en figure 6.2a. Elle se compose en 5 étapes. La première étape consiste en une première couche correspondant aux deux *input* du réseau, c'est-à-dire à une représentation vectorielle pour chaque modalité du document (dans notre cas textuelle et visuelle). Une seconde couche correspond à l'encodage de chacune des deux modalités de façon indépendante selon le principe des *autoencoders*. Une couche centrale réunit ensuite les deux modalités afin d'obtenir une unique représentation du document. Les deux couches suivantes sont similaires aux deux premières, à savoir une reséparation des deux modalités avant un *output* devant être identique à l'*input*.

L'originalité du modèle BiDNN par rapport aux autres autoencoders mutlimodaux est qu'il scinde en deux la phase d'apprentissage : une pour chaque modalité. La crossmodalité est conservée par le fait que, sur les deux réseaux appris, les poids liés aux couches centrales sont symétriques, comme illustré sur la figure 6.2b. D'un point de vue pratique, les variables correspondant aux poids sont les mêmes d'un côté et de l'autre. L'apprentissage s'effectue en utilisant alternativement un passage de la modalité textuelle vers la modalité visuelle et inversement. Une fois le modèle appris, il suffit de calculer les *encodings* de chacune des deux modalités et de les concaténer pour obtenir une représentation bimodale du document.

Formellement, le réseau peut être décrit ainsi. Soit $\mathbf{h_i^{(j)}}$ l'activation d'une couche cachée à la profondeur j dans le réseau i ($i = 1, 2$, un pour chaque modalité), $\mathbf{x_i}$ le vecteur

2. Ce système a été imaginé et développé par Vedran Vucotić, membre de l'équipe LinkMedia. Notre contribution a consisté à fournir des conseils sur les moyens d'évaluer son efficacité ainsi qu'à sa mise en œuvre dans la campagne TRECVid 2016 et l'évaluation de sa diversité.



(a) Architecture d'un autoencoder bimodal classique.

(b) Architecture d'un autoencoder bidirectionnel (BiDNN).

FIGURE 6.2 – Deux architectures d'autoencoders bimodaux (Vukotić et al., 2016).

caractéristique de la modalité i et $\mathbf{y_i}$ l'*output* du réseau pour la modalité i ; les réseaux peuvent être définis par leurs matrices de poids $\mathbf{W_i^{(j)}}$ et leurs vecteurs de biais $\mathbf{b_i^{(j)}}$ pour chaque couche j, et utilisent la fonction d'activation f, tanh dans notre cas. L'architecture peut alors être définie par :

$$\mathbf{h_i^{(1)}} = f(\mathbf{W_i^{(1)}} \times \mathbf{x_i} + \mathbf{b_i^{(1)}}) \qquad i = 1,2 \tag{6.3}$$

$$\mathbf{h_1^{(2)}} = f(\mathbf{W^{(2)}} \times \mathbf{h_1^{(1)}} + \mathbf{b_1^{(2)}}) \tag{6.4}$$

$$\mathbf{h_1^{(3)}} = f(\mathbf{W^{(3)}} \times \mathbf{h_1^{(2)}} + \mathbf{b_1^{(3)}}) \tag{6.5}$$

$$\mathbf{h_2^{(2)}} = f(\mathbf{W^{(3)}}^{\mathbf{T}} \times \mathbf{h_2^{(1)}} + \mathbf{b_2^{(2)}}) \tag{6.6}$$

$$\mathbf{h_2^{(3)}} = f(\mathbf{W^{(2)}}^{\mathbf{T}} \times \mathbf{h_2^{(2)}} + \mathbf{b_2^{(3)}}) \tag{6.7}$$

$$\mathbf{o_i} = f(\mathbf{W_i^{(4)}} \times \mathbf{h_i^{(3)}} + \mathbf{b_i^{(4)}}) \qquad i = 1,2 \tag{6.8}$$

Les matrices de poids $\mathbf{W^{(2)}}$ et $\mathbf{W^{(3)}}$ sont utilisées deux fois (équations 6.4, 6.7 et équations 6.5, 6.6) du fait qu'il s'agisse des mêmes poids. L'apprentissage s'effectue en minimisant la moyenne de l'erreur au carré (*mean squared error*) de $(\mathbf{o_1}, \mathbf{x_2})$ et $(\mathbf{o_2}, \mathbf{x_1})$, ce qui permet de minimiser l'erreur dans les deux directions et de créer un espace de représentation commun dans la couche centrale.

Lors de nos expérimentations, les tailles retenues pour les couches cachées sont de 200-100-200, des tailles plus grandes n'apportant pas de gains de performance notables. La représentation utilisée pour la modalité textuelle en *input* du réseau est un Word2Vec (Mikolov et al., 2013) moyenné pour la modalité textuelle (vecteurs de taille 100). Cette représentation a été utilisée dans plusieurs autres travaux avec succès (Yang et al., 2016). La librairie Python Gensim a été utilisée afin d'obtenir ces représentations [3]. La modalité visuelle a été représentée grâce aux concepts visuels fournis par EUROCOM

3. www.github.com/RaRe-Technologies/gensim

dans le cas de l'évaluation décrite en section 6.3.2, et une représentation VGG-19 (vecteurs de taille 4096) lors de notre participation à la campagne TRECVid 2016 (Bois et al., 2016). L'entraînement du BiDNN a été réalisé à l'aide d'un gradient stochastique avec une quantité de mouvement de Nesterov (*Nesterov momentum*) et un *dropout* de 20 %. L'implémentation, réalisée avec Lasagne [4], est disponible en ligne [5].

6.3 Évaluations

L'évaluation des deux systèmes précédents est complexe car leur objectif est double : offrir une forte pertinence des liens créés, tout en assurant une diversité importante, voire de la sérendipité. Nous proposons donc ici une double évaluation, une première fondée sur la pertinence des liens et évaluée dans le cadre des campagnes TRECVid, et une seconde fondée sur la diversité offerte et réalisée à l'aide d'études utilisateurs menées par nos soins. Nous démontrons la pertinence des deux modèles présentés plus tôt selon ces deux mesures.

6.3.1 Scores de pertinence

Nous avons expérimenté le CrossLDA et le BiDNN lors de deux campagnes TRECVid, à savoir l'édition 2015 (Simon et al., 2015) pour le premier et l'édition 2016 (Bois et al., 2016) pour le second. Bien que les ancres, et plus encore les corpus soient différents dans ces deux éditions, la comparaison des résultats des systèmes reste intéressante. Lors de chacune des éditions de TRECVid, la pertinence a été évaluée par des annotateurs humains via la plateforme Amazon Mechanical Turk, à raison d'une unique évaluation par paire (ancre-cible). Des expérimentations supplémentaires postérieures de la part des organisateurs montrent que cette unique évaluation a des défauts de fiabilité, une seconde évaluation par paire menant à un désaccord dans plus d'un cas sur 5 (Ordelman et al., 2015).

Lors de l'édition 2015, le CrossLDA a été utilisé selon deux configurations différentes. Une première (RUN-1) a consisté à aller de la modalité textuelle vers la modalité visuelle, en utilisant pour les ancres la représentation textuelle fournie par le CrossLDA, et en comparant cette représentation aux représentations visuelles des cibles potentielles. Une seconde (RUN-2) a consisté en l'approche inverse, et donc à utiliser une représentation visuelle pour l'ancre et une représentation textuelle pour les cibles. Une *baseline* correspondant à une similarité de surface fondée uniquement sur la modalité visuelle a également été utilisée à des fins de comparaison. La table 6.2 récapitule les scores de précision au rang 5 et au rang 10 de ces trois méthodes. Si les scores de la RUN-1 laissent penser qu'une erreur a conduit à des résultats très faibles, on constate que l'utilisation du CrossLDA allant de la modalité visuelle à la modalité textuelle améliore les résultats par rapport à une approche de surface.

Lors de l'édition 2016, le BiDNN a été utilisé. Deux *baseline*s fondées sur une seule des deux modalités ont été utilisées à des fins de comparaison. Ces *baseline*s correspondent aux représentations vectorielles monomodales fournies en entrée du modèles BiDNN. Il s'agit donc d'un Word2Vec moyenné pour la *baseline* audio, et d'un vecteur VGG-19 pour la version visuelle. La table 6.3 récapitule les scores de ces différents systèmes. On

4. www.github.com/Lasagne/Lasagne
5. www.github.com/v-v/BiDNN

	baseline	RUN-1	RUN-2
prec@5	0.21	0.016	**0.26**
prec@10	0.20	0.017	**0.22**

TABLE 6.2 – Résultats des approches CrossLDA à TRECVid 2015.

	Audio	Visuel	BiDNN
prec@5	0.40	0.45	**0.52**

TABLE 6.3 – Résultat du BiDNN et de deux *baseline*s monomodales.

remarque que la représentation monomodale visuelle est meilleure que la représentation audio. Le BiDNN, crossmodal, améliore sensiblement la précision au rang 5. La précision au rang 10 n'ayant pas été calculée lors de l'édition 2016, nous n'en fournissons pas les scores.

6.3.2 Évaluation humaine de la diversité

La pertinence seule ne suffit pas à évaluer l'intérêt d'un système d'hyperliage. Comme mentionné dans la section 6.1, la diversité est un second enjeu majeur pour ces outils. Afin d'évaluer la diversité des approches BiDNN et CrossLDA, nous avons appliqué ces deux méthodes sur le corpus TRECVid 2015, et conduit une expérimentation à l'aide d'évaluateurs humains. La version CrossLDA évaluée est celle ayant fourni les meilleurs résultats lors de la campagne, à savoir le modèle visuel vers texte. Les cibles proposées par ce système lors de l'édition 2015 ont été réutilisées dans les expérimentations qui suivent. Étant donné que le BiDNN n'avait pas participé à la campagne 2015, nous avons réentrainé le système sur le corpus correspondant, et appliqué un réordonnancement des cibles proposées par l'ensemble des participants à la tâche. Utiliser les cibles proposées par les participants de cette édition nous permet en effet de réutiliser les évaluations de la pertinence de celles-ci. Enfin, une nouvelle *baseline*, fondée sur la modalité langagière a été construite. Celle-ci consiste à représenter chaque segment par un vecteur des mots prononcés pondérés par la mesure tf-idf. Un cosinus est ensuite appliqué entre les paires de segments afin d'obtenir un score de similarité servant à l'ordonnancement des cibles. Pour cette *baseline*, et de façon similaire à l'approche BiDNN, seules les cibles proposées par les participants de 2015 ont été utilisées.

Ces trois méthodes construites et une liste ordonnée des segments établie, nous avons pu évaluer, grâce aux résultats de l'édition 2015, la proportion de cibles pertinentes. Nous avons ensuite sélectionné un sous-ensemble d'ancres pour lesquels les trois systèmes étudiés obtenaient les meilleures performances. Ce choix a été réalisé afin de d'éviter les cas où certains systèmes n'auraient pas suffisamment de cibles pertinentes à proposer. Ainsi, sur les 100 ancres proposées initialement, nous en avons retenu 16 pour lesquelles les scores des trois méthodes donnaient les meilleurs résultats. La table 6.4 récapitule les scores de pertinence des trois systèmes sur les 100 ancres et sur l'échantillon de 16 ancres. On note que la *baseline* et le BiDNN obtiennent des scores similaires, tandis que le CrossLDA obtient des résultats inférieurs. La très bonne performance de la *baseline*, pourtant très simple, montre la présence de nombreuses cibles jugées pertinentes très similaires à leurs ancres respectives d'un point de vue lexicographique.

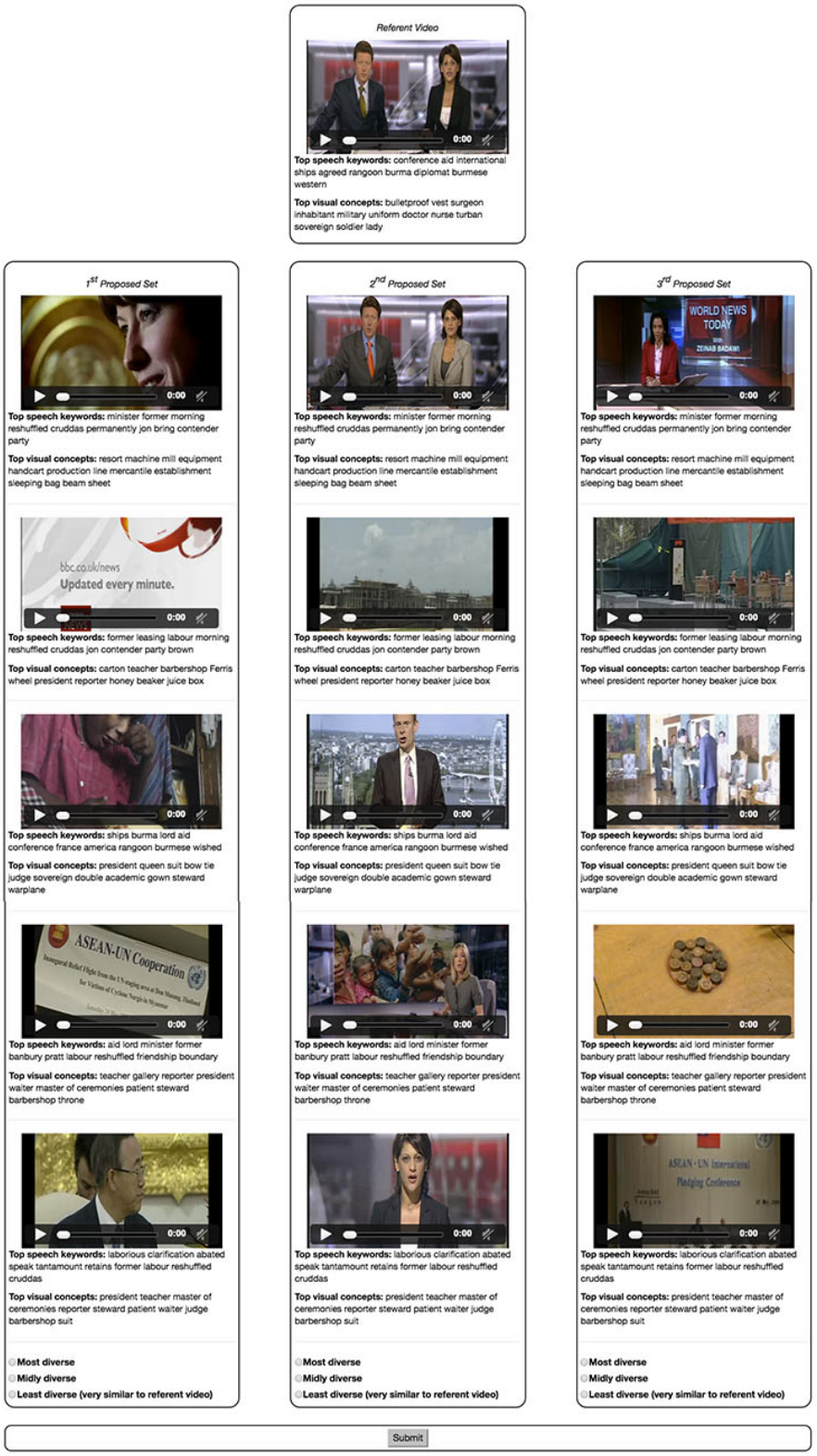


FIGURE 6.3 – Formulaire d'évaluation de la diversité.

Ancres	baseline	BiDNN	CrossLDA
100	0.59	0.57	0.24
16	0.80	0.80	0.40

TABLE 6.4 – Précision au rang 10 sur la tâche de réordonnancement.

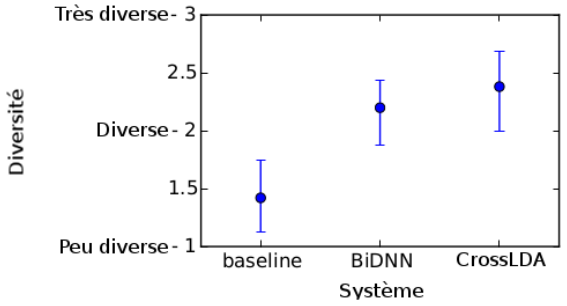


FIGURE 6.4 – Diversité moyenne des systèmes telle que perçue par les utilisateurs.

Une fois ces bases posées, l'évaluation de la diversité des trois systèmes a pu être menée. Nous avons commencé par sélectionner, pour chacune des 16 ancres, les 5 premières cibles jugées pertinentes par chaque système. Nous obtenons donc, pour chaque ancre, un jeu de 15 cibles pertinentes, réparties en trois groupes, un groupe par système. Nous présentons ensuite aux évaluateurs l'ancre, ainsi que les 15 cibles organisées en trois colonnes, une par système. Ceux-ci doivent alors décider quelle colonne propose le jeu de cibles le plus diversifié, et laquelle propose le jeu le moins diversifié. Après son évaluation, une nouvelle ancre est proposée à l'évaluateur. L'ordre des colonnes est aléatoire, et aucune indication ne permet à l'évaluateur de savoir quel système est associé à chaque colonne. Afin d'aider l'évaluateur dans sa tâche, quelques mots-clés sont extraits automatiquement des segments (ancres et cibles) à l'aide d'une pondération tf-idf (Hasan et Ng, 2010). Selon la même méthode, des concepts-clés, correspondant aux concepts visuels les plus importants présents dans les segments, sont extraits et présentés aux évaluateurs. L'interface d'évaluation, telle qu'elle a été utilisée, est illustrée dans la figure 6.3.

25 évaluateurs, issus principalement du monde académique mais pour la plupart sans connaissance de la tâche d'hyperliage, ont participé à l'étude. Au total, 176 votes ont été enregistrés, soit en moyenne 11 votes par ancre. La figure 6.4 présente les rangs moyens, ainsi que les intervalles de confiance pour les trois systèmes. Le CrossLDA obtient le meilleur résultat avec un rang moyen de 2,38, suivi par le BiDNN avec un score de 2,20. La *baseline* est très largement en dessous des deux autres système avec un rang moyen de 1,42. Une différence significative ($\rho < 0.01$ test de Student) est observée entre chaque paire de systèmes. Ainsi, bien que la *baseline* et le BiDNN aient des scores de pertinence très similaires, le BiDNN se démarque par une diversité de résultats largement supérieure, ce qui en fait un système probablement plus intéressant pour une tâche d'hyperliage. La différence significative entre le CrossLDA et le BiDNN est plus complexe à analyser. En effet, selon Good et al. (1999), la diversité évolue comme une fonction inverse de la pertinence. Il est donc difficile de conclure étant donné que le CrossLDA, s'il est significativement plus divers, est néanmoins significativement moins pertinent.

	transcriptions			concepts		
	n_u	$\overline{d}_a$	$\overline{d}_i$	n_u	$\overline{d}_a$	$\overline{d}_i$
baseline	29.8	0.51	0.61	35.6	0.61	0.71
BiDNN	40.8	0.20	0.12	46.7	0.42	0.31
CrossLDA	40.0	0.25	0.16	38.0	0.48	0.41

TABLE 6.5 – Évaluation automatique de la diversité des cibles.

6.3.3 Mesures automatiques pour la diversité

En complément de l'étude utilisateurs décrite précédemment, nous avons analysé les scores de similarité inter-cibles. Ainsi, pour chaque groupe de cibles pertinentes présenté dans l'étude utilisateur, nous avons comptabilité le nombre moyen de mots-clés uniques et de concepts-clés uniques, noté n_u. Un score de 10 indique donc que toutes les cibles étaient décrites par les mêmes mots-clés (resp. concepts-clés) tandis qu'un score de 50 indique que chaque cible proposait des mots-clés (resp. concepts-clés) uniques. Nous avons également mesuré la similarité de surface moyenne entre chaque ancre et ses cibles (notée $\overline{d}_a$). Cette mesure de similarité a été réalisée l'aide d'une représentation tf-idf et d'une mesure cosinus. Enfin, nous avons mesuré de la même manière la similarité de surface moyenne entre les cibles d'une même ancre (notée $\overline{d}_i$).

La table 6.5 rapporte les résultats de ces mesures. On remarque que ces mesures automatiques sont en partie en corrélation avec les évaluations manuelles décrites dans la section précédente. En effet, les méthodes CrossLDA et BiDNN y sont significativement meilleures que la *baseline* (n_u supérieure et $\overline{d}_a$, $\overline{d}_i$ inférieures sur les deux modalités). En revanche, selon ces mesures automatiques, BiDNN et CrossLDA se montrent très proches, avec un avantage léger pour BiDNN, infirmé par l'étude utilisateur présentée plus tôt. Ces mesures automatiques simples permettent donc de comparer des systèmes dont les différences de performance en terme de diversité sont suffisamment importantes, mais ne semblent pas capables de remplacer l'étude humaine pour des systèmes aux performances proches. Cette étude comparative sur la diversité a fait l'objet d'une publication à l'international (Bois et al., 2017c).

6.4 Orienter la diversité : le LDA hiérarchique

La recherche de diversité est un enjeu majeur lorsque le besoin d'information n'est pas explicite, ou qu'il est ambigu. Néanmoins, lorsque le besoin est précis et clairement spécifié, offrir une large diversité revient à prendre un risque sur la pertinence des liens proposés. Pour ces raisons, la possibilité de contrôler la diversité au sein d'un même algorithme devient une problématique intéressante. Dans cette section, nous décrivons un modèle d'hyperliage permettant le contrôle de la diversité des liens : le LDA hiérarchique.

6.4.1 Méthode

L'une des difficultés posées par le CrossLDA, vu en section 6.2.2, est de fixer le nombre de *topics* latents à l'aide du paramètre K. Ceci est valable également pour le LDA classique, monomodal. Une valeur de K trop élevée multipliera les *topics*, les documents obtenant des distributions semblables sur ces *topics* risquant alors d'être tous très similaires,

et donc peu diversifiés. À l'inverse, une valeur de K trop faible ne permettra pas de distinguer des thématiques pourtant sémantiquement éloignées, risquant de créer des liens entre des documents peu semblables.

Une solution à ce problème consiste à utiliser conjointement plusieurs tailles de *topics*, allant du général (K faible) au spécifique (K élevé). En combinant ces différents niveaux, on peut alors choisir à quel niveau de spécificité créer les liens. Plus précisément, nous proposons trois façons de combiner les différents niveaux de spécificité : une combinaison simple, dans laquelle les *topics* appris pour des valeurs de K différentes sont indépendants, une structure en arbres naïve, permettant d'expliciter le lien thématique qui unit deux documents, et une structure en arbres contrôlée, permettant d'assurer de meilleures propriétés topologiques à la hiérarchie créée [6].

Pour chacune de ces structures, une étape préliminaire consiste à apprendre les *topics* pour plusieurs valeurs de K sur la collection. Dans nos expérimentations, nous choisissons $K \in 50, 150, 300, 700, 1500$ pour les méthodes de combinaison simple et de structure en arbres naïve et $K \in 50, 150, 300, 700$ pour la structure en arbres naïve et contrôlée.

La méthode de la combinaison simple consiste à considérer tous les niveaux appris, et à créer une nouvelle représentation pour chaque document issue de la combinaison des distributions pour chaque niveau K. Ainsi, la formule du LDA classique déjà explicitée en section 6.2.2 :

$$p(d|z_j) = (\prod_{i=1}^{n_d} p(w_i, z_j))^{\frac{1}{n_d}} \tag{6.9}$$

peut être utilisée afin d'obtenir une représentation pour chaque niveau de spécificité l :

$$p(d) = [p(d|z_1^l), p(d|z_2^l), p(d|z_3^l), \ldots, p(d|z_{K_l}^l)] \tag{6.10}$$

Ces différents niveaux peuvent ensuite être réutilisés lors de la comparaison entre les représentations de deux documents d_1 et d_2 par la formule :

$$S(d_1, d_2) = -\sum_l \alpha_l \log(d_1, d_2) \tag{6.11}$$

où α_l permet de pondérer chacun des niveaux afin de donner davantage d'importance aux thématiques spécifiques (K élevé) ou générales (K faible).

La méthode de structure en arbres naïve consiste à tirer partie des différents niveaux thématiques afin de les organiser en hiérarchie. Ainsi, à chacun des *topics* à un niveau général l (*e.g.*, $K = 50$), on fera correspondre un ou plusieurs *topics* au niveau plus spécifique $l + 1$ (*e.g.*, $K = 150$). Cette correspondance est établie grâce à une similarité logarithmique entre les distributions de mots de chaque paire de *topics*. Ainsi, la distribution du topic z_i^l sera comparée aux distributions des *topics* du niveau inférieur $z_j^l + 1 \forall j$. La paire $(z_i^l, z_j^l + 1)$ obtenant le score de similarité le plus élevé mènera à la création d'un lien entre z_i^l et $z_j^l + 1$. Ce processus est répété pour chaque niveau de spécificité ($K = 150$ avec $K = 300$, $K = 300$ avec $K = 700$ et $K = 700$ avec $K = 1500$). On obtient ainsi une structure arborée telle que montrée figure 6.5. On peut alors représenter un document d non plus par l'ensemble des *topics* à chaque niveau, mais par la série de *topics* t^d telle que $t^d = [t_1^d, t_2^d, \ldots, t_l^d]$ avec $t_1^d = max(z_i^{l_{max}})$ et $t_{n+1}^d = parent(t_n^d)$. En d'autres termes, après

6. La création d'une structure hiérarchique LDA fondée sur différentes valeurs de K a été proposée par Anca Simon, membre de l'équipe LinkMedia. Notre contribution a consisté a proposer une amélioration théorique de ce modèle dénommée ici structure arborée contrôlée et à l'implémenter.

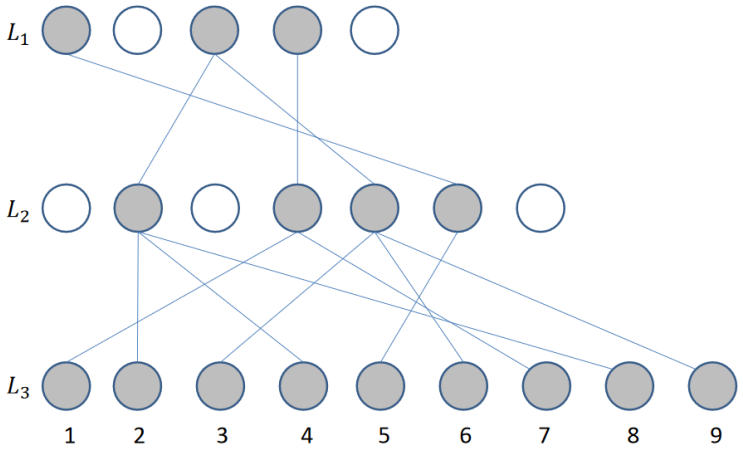


FIGURE 6.5 – Structure arborée naïve, extraite de Simon (2015).

avoir trouvé le *topic* spécifique le plus probable pour d, on remonte la structure arborée depuis ce *topic* jusqu'au *topic* le plus général.

Cette série de *topics*, accompagnée des scores de probabilités de d leur correspondant, sera alors utilisée comme représentation. On peut alors réutiliser la formule pour comparer deux documents :

$$S(d_1, d_2) = -\sum_l \alpha_l \log(t^{d_1}, t^{d_2}) \tag{6.12}$$

Si cette structuration de la hiérarchie de *topics* permet de maximiser la similarité entre un *topic* enfant et son *topic* parent, elle mène invariablement à une structure déséquilibrée dans laquelle certains *topics* génériques ont beaucoup d'enfants tandis que d'autres se retrouvent sans aucun descendant. Ce phénomène est illustré en figure 6.5 où l'on constate que si tous les *topics* les plus spécifiques ont un parent, certains *topics* des niveaux l_1 et l_2 ne disposent d'aucun enfant. Ces *topics* oubliés, inaccessibles, n'entreront par conséquence jamais dans les représentations t^d des documents. Afin d'éviter ce phénomène, nous proposons d'avoir recours à la programmation entière linéaire (*Integer Linear Programming* ou ILP), une méthode fondée sur la définition de contraintes, afin d'équilibrer la structure. Ainsi, en suivant la formulation suivante, on oblige chaque *topic* à avoir au moins 2 enfants, tout en gardant comme objectif de maximiser la similarité parent-enfant.

$$\begin{aligned} \text{Maximiser} : & \sum_{i,j} \text{sim}(z_i^l, z_j^{l+1}) \text{link}(z_i^l, z_j^{l+1}) \\ \text{Tel que} : & \sum_i \text{link}(i,j) = 1 \qquad \forall j \\ Et : & \sum_j \text{link}(i,j) \leq 2 \qquad \forall i \end{aligned} \tag{6.13}$$

Dans cette formulation, la première ligne correspond à la fonction à maximiser, soit ici la similarité parent-enfant ; la deuxième ligne correspond à la contrainte de n'avoir qu'un unique parent pour chaque enfant ; la troisième ligne correspond à la contrainte que tout parent doit avoir au moins deux enfants. Lors de nos expérimentations, le *solver* glpk [7] a été utilisé pour résoudre le système d'équations, via la librairie Python PuLP [8]. On peut

7. www.gnu.org/software/glpk/
8. www.github.com/coin-or/pulp

meilleur *topic* spécifique	*topic* frère	*topic* général
city	great	people
people	city	world
place	empire	war
good	roman	city
countryside	world	british
heart	christian	britain
centre	building	life
nation	living	great
visit	light	work
capital	modern	history

TABLE 6.6 – Exemple issu d'une structure parent-enfant.

utiliser cette structuration de la même façon que la structure arborée naïve, c'est-à-dire l'utiliser afin d'obtenir une représentation t^d des documents. Les deux méthodes de structuration permettent également d'illustrer la nature de la relation entre deux documents en visionnant les *topics* qui les unissent. On peut aussi choisir de s'écarter légèrement de t^d pour la sélection de la cible afin d'augmenter la diversité. On peut en effet sélectionner la feuille sœur de t^d au niveau le plus spécifique, tout en conservant le reste de la représentation commune pour les deux documents. Ainsi, l'ancre et la cible auront chacune une représentation t^d utilisant les mêmes *topics*, à l'exception du niveau le plus spécifique. La table 6.6 montre un exemple issu d'un hyperliage jugé pertinent lors de la tâche Hyperlinking MediaEval 2013, identique à la campagne MediaEval 2014 à l'exception du jeu d'ancres évalué. Cette option, ajoutée à celle consistant à pondérer différemment les *topics* génériques et les *topics* spécifiques le long de la représentation t^d, permettent un contrôle accru sur la diversité souhaitée pour l'utilisateur.

6.4.2 Évaluation

Les trois approches décrites précédemment ont été évaluées sur les données issues des campagnes MediaEval 2013 et 2014. Il s'agit, comme pour la campagne TRECVid 2015, de 3 mois de vidéos provenant de la BBC. Ces systèmes n'ayant pas pu être évalués directement lors des différentes campagnes, nous avons récupéré les cibles proposées par les participants, et transformé la tâche en une tâche de réordonnancement des cibles potentielles. L'objectif consiste alors à donner un score élevé aux cibles jugées pertinentes, et un score faible à celles évaluées négativement lors de la campagne.

La table 6.7 donne les résultats des trois méthodes ainsi que deux *baseline*s. La première *baseline* correspond à une mesure de similarité directe entre ancres et cibles, fondée sur une représentation tf-idf des transcriptions des segments vidéos. Une mesure cosinus est ensuite appliquée sur ces représentations afin d'obtenir des scores de similarité. Cette *baseline* a pour objectif de trouver des segments très similaires, laissant peu de place à la diversité. La seconde *baseline* correspond à l'exploitation d'un modèle LDA classique, utilisant une unique valeur K. Nous rapportons ici les résultats pour $K = 150$, qui a obtenu les meilleurs scores sur cette tâche. Les scores des trois méthodes décrites précédemment sont ensuite répertoriés, $\text{Comb}_{=}$, $\text{Comb}_{<}$, $\text{Comb}_{>}$ correspondant respectivement à la combinaison simple avec sans pondération (chaque niveau dispose du même poids), avec une pondération croissante (poids plus important pour les thématiques spé-

	2013			2014		
	@5	@10	@20	@5	@10	@20
Direct	0.71	0.66	0.62	0.41	0.41	0.38
T150	0.67	0.64	0.58	0.45	0.4	0.35
$\text{Comb}_{=}$	0.7	0.67	0.63	0.34	0.33	0.31
$\text{Comb}_{<}$	0.68	0.66	0.62	0.31	0.33	0.32
$\text{Comb}_{>}$	0.71	0.68	0.63	0.35	0.35	0.33
Arb_1	0.54	0.49	0.43	0.43	0.38	0.35
Arb_2	0.44	0.44	0.39	0.43	0.43	0.37
Arb_C	0.4	0.39	0.37	0.43	0.44	0.4

TABLE 6.7 – Performances en terme de pertinence pour les trois stratégies de combinaisons thématiques.

cifiques) et une pondération décroissante (poids plus faible pour les thématiques spécifiques). Ces trois résultats utilisent des valeurs de $K \in 50, 150, 300, 700, 1500$. Arb_1 et Arb_2 correspondent à la structure d'arbres naïve selon deux variations de hiérarchies, la première sur la plage de valeurs $K \in 50, 150, 300, 700, 1500$ et la seconde sur la plage $K \in 50, 150, 300, 700$. Enfin, Arb_C correspond à la structure d'arbres contrôlée sur la plage $K \in 50, 150, 300, 700$.

On remarque que la combinaison de plusieurs niveaux utilisées dans Comb améliore sensiblement les scores de l'unique niveau utilisé dans T150 pour l'année 2013. Ce n'est en revanche pas le cas sur l'année 2014. Les structures arborées offrent quant à elles des résultats comparables aux *baseline*s sur l'année 2014. Globalement, ces résultats justifient l'utilisation de méthodes fondées sur la détection de *topics* et leur exploitation sous forme de hiérarchie. En effet, les résultats ne semblent pas dégradés par l'utilisation de ces approches, alors même que ces dernières permettent une diversité des liens accrue, ainsi qu'un meilleur contrôle de la diversité. Les résultats de ces expérimentations ont été publiés dans Şimon et al. (2015).

Conclusion

La diversité des systèmes d'hyperliage est complexe à évaluer, et généralement peu ou pas récompensée lors de campagnes d'évaluation où seule la pertinence est considérée. L'étude utilisateur que nous avons proposée, si elle se révèle relativement coûteuse en temps d'annotation, nous paraît indispensable à l'évaluation objective de l'intérêt d'un système d'hyperliage. Il semble que les indicateurs automatiques que nous proposons permettent de différencier des systèmes offrant une diversité importante de systèmes peu divers, diminuant d'autant le besoin d'avoir recours à des évaluateurs humains, nécessaires pour comparer les systèmes aux performances plus proches. L'explication de la sémantique des liens créés, par exemple en montrant que deux documents partagent une même thématique générale, ainsi que le contrôle de la diversité, sont deux enjeux majeurs auxquels le LDA hiérarchique apporte une première réponse. Dans le chapitre suivant, nous proposons d'aller plus loin dans l'explication des liens, en réalisant un typage explicite permettant aux utilisateurs de comprendre la sémantique reliant deux documents.

Troisième partie

Enrichissement par typage d'hyperliens pour une navigation éclairée

Chapitre 7

Typologie de liens : description et construction

Introduction

La construction de liens pertinents et diversifiés, regroupés au sein d'un hypergraphe, a pour but de permettre une navigation aisée d'une collection, lors de laquelle l'ensemble des documents sont accessibles. Néanmoins, d'un point de vue pratique, la navigation s'exerce par le biais d'un ensemble de suggestions proposées pour chacun des documents. Ainsi, l'utilisateur se retrouve confronté à des documents, et doit faire l'effort cognitif de comprendre la relation qui existe entre le document qu'il visualise et chacun dee ceux qui lui sont proposés à cette étape. Or, la diversité même des liens offerts implique une diversité des relations entre documents plus importante, compliquant d'autant plus l'effort demandé.

Pour répondre à cette problématique, nous proposons de catégoriser les liens existant entre paires de documents à l'aide d'une typologie développée pour le corpus LIMAH. L'objectif consiste à expliciter la relation entre chaque paire de documents, afin que l'utilisateur sache, avant même de cliquer sur un lien, ce qu'il peut s'attendre à y trouver.

Dans ce chapitre, nous décrivons la typologie construite pour le typage des liens au sein du corpus LIMAH, donnons des exemples d'instanciation de ce typage, et explicitons les méthodes employées pour effectuer un typage automatique. Celui-ci est fondé sur des heuristiques simples, paramétrées dans le but d'obtenir une bonne répartition des types de liens. Une discussion sur la subjectivité du typage est également conduite.

7.1 Typologie

Une étape préliminaire au typage explicite des liens entre documents consiste à étudier les types pertinents pour les usages visés. Nous nous appuyons pour cela sur l'étude utilisateur menée en section 1.2.3 ainsi que sur une analyse du corpus LIMAH que nous avons élaboré. Dans cette section, nous présentons un état de l'art des différentes typologies existantes, avant d'en proposer une adaptée à notre corpus et à notre cas d'utilisation. Nous illustrons cette typologie au travers d'exemples et discutons en détail de la

subjectivité de l'assignation d'un lien.

7.1.1 État de l'art

La création de liens entre documents, ou fragments de document, a été explorée par différentes communautés. Celle du traitement automatique des langues s'est principalement intéressée, au travers de corpus journalistiques, à lier des événements entre eux, le plus souvent via deux types de liens : un lien de similarité (les deux documents présentent le même événement) et un lien de causalité temporelle (un événement en provoque un second) (Nallapati et al., 2004; Rennison, 1994; Muller et Tannier, 2004). Ces relations temporelles, plus largement décrites dans la section 2.2, facilitent le parcours d'une collection de documents en proposant une navigation chronologique permettant de recomposer l'évolution d'une série d'événements. Il nous semble néanmoins que l'utilisation de ces deux seuls types de liens est peu adaptée à un corpus multimédia, dans lequel de nombreux événements sont décrits simultanément par différentes sources, commentés sur les réseaux sociaux et repris par des blogueurs. Un typage plus fin, et davantage en phase avec la notion d'information qu'avec celle d'événement, nous paraît donc nécessaire. Une seconde approche explorée par la communauté du traitement des langues consiste à relier des thèmes (*topics*) sous-jacents aux documents. Une fois encore, le but principal consiste à regrouper ces thèmes et à proposer un parcours chronologique de ceux-ci (Ide et al., 2004), avec les mêmes limites que celles décrites précédemment.

La communauté du multimédia s'est également intéressée à la création automatique de liens entre documents. Le plus souvent, ces liens ne sont pas typés et ne servent qu'à mettre en lumière une relation non explicite entre deux documents, comme dans le cas de l'hyperliage largement décrit plus tôt (Eskevich et al., 2012). Ces liens non explicites se révèlent particulièrement utiles pour de petites collections, ou pour le développement de moteurs de recommandation. Cette absence de typage limite néanmoins les usages possibles et rend difficile une navigation éclairée. D'autres travaux dans ce même domaine mettent d'ailleurs en avant le besoin d'un typage pour faciliter le parcours de grandes collections (Cleary et Bareiss, 1996). Cette dernière étude expose une partie de la typologie qu'elle utilise où les 8 types les plus fréquents sont décrits sous forme de questions. Nous y trouvons des types classiques du domaine journalistique tels que les relations de causalité, mais aussi des liens de type « conseils : comment puis-je capitaliser sur cette situation ? » qui sont difficilement instanciables sur un corpus d'actualités.

La communauté des sciences de l'information et de la communication s'est aussi penchée sur le rôle des liens hypertextes. L'un de ces travaux a notamment influencé la typologie que nous proposons (Ertzscheid, 2002). Cette étude, bien que très générique et se plaçant dans un contexte éloigné du domaine journalistique, met en avant plusieurs grandes catégories de liens dont nous nous inspirons (*e.g.* une relation de récurrence, peu abordée dans les autres travaux, mais qui nous paraît pertinente dès lors que le corpus considéré est volumineux).

7.1.2 Description de la typologie

Nous proposons trois grandes catégories de liens, dont deux sont divisées en sous-catégories. Pour chacun des liens présentés, un lien inverse existe de telle sorte que tout fragment de document lié à un autre est à la fois source et cible d'un lien. Cette relation double peut se caractériser par des liens non orientés (*e.g.* un lien de type quasi duplicat

est non orienté) ou par des liens duaux (*e.g.* le lien dual du développement est le résumé). Les types proposés ne sont pas exclusifs, un lien entre deux documents pouvant disposer de plusieurs types (*cf.* section 7.1.3). Les trois catégories retenues sont :

la récurrence : répétition d'une information. Le contenu est similaire mais peut être présenté de diverses manières, indépendamment de la modalité utilisée ;

l'extension : enrichissement d'une information. L'extension peut correspondre à un enrichissement en volume, avec un contenu plus large, ou bien à une extension temporelle correspondant à un suivi d'information ;

la réaction : l'information est commentée par un nouvel intervenant.

La récurrence est la relation la plus fréquemment rencontrée. Elle peut être envisagée sous trois formes :

le quasi duplicat : l'information est répétée, de façon similaire, sans ajout ou suppression notable ;

la citation : une référence à une information délivrée précédemment est incluse ;

la parodie : une information est reprise et détournée.

Nous considérons la parodie comme une forme de récurrence car elle reprend une information identique et change son traitement à des fins de divertissement. L'information traitée reste néanmoins la même.

L'extension enrichit une information en la développant ou en exhibant un lien temporel avec une autre information. Elle peut donc se préciser selon les deux sous-catégories suivantes :

le développement : l'information est développée, son contenu est plus important ;

la postériorité : une relation de suivi temporel est exhibée entre les deux informations.

La réaction concerne l'ensemble des commentaires qui peuvent être apportés sur une information, que ceux-ci aient lieu dans un milieu contrôlé (*e.g.* diffusion de la réponse d'un homme politique à une critique adverse) ou libre (*e.g.* réaction d'un internaute sur Twitter). Nous choisissons de ne pas offrir de sous-catégories à la réaction, bien qu'il soit possible d'utiliser les typologies existantes en analyse d'opinion pour affiner ce type (Ekman, 1992).

La typologie proposée est donc issue à la fois d'un réagencement de types couramment utilisés, ainsi que de types rarement utilisés, mais pertinents dans le cadre d'un corpus multimodal diversifié. Elle reprend donc les relations classiques d'antériorité/-postériorité, qui permettent de suivre une information d'un point de vue temporel, ou bien de source/citation, largement étudiées dans le cadre de corpus scientifiques (Nanba et al., 2011; Thelwall, 2003), mais aussi des relations moins souvent exploitées telles que la parodie ou le quasi duplicat.

La figure 7.1 présente la typologie proposée par cette étude. Elle indique la nature des relations duales lorsqu'elles existent. Lorsqu'il n'y a pas de dualité, le lien est considéré comme non orienté. Ainsi, un lien d'antériorité entraîne nécessairement un lien inverse de postériorité, ou un lien de développement correspond toujours à un lien de résumé. Cette typologie nous paraît couvrir une très large majorité des liens possibles et permet d'envisager de nouveau moyens de parcourir une grande collection de documents.

La typologie que nous décrivons dans cette section, et qui a fait l'objet d'une publication (Bois et al., 2015), a été conçue avant l'obtention finale du corpus LIMAH décrit dans la section 3.2, grâce à des échantillons. Nous proposons dans cette typologie deux liens qui n'auront finalement pas été utilisés dans la suite des expérimentations : la sour-

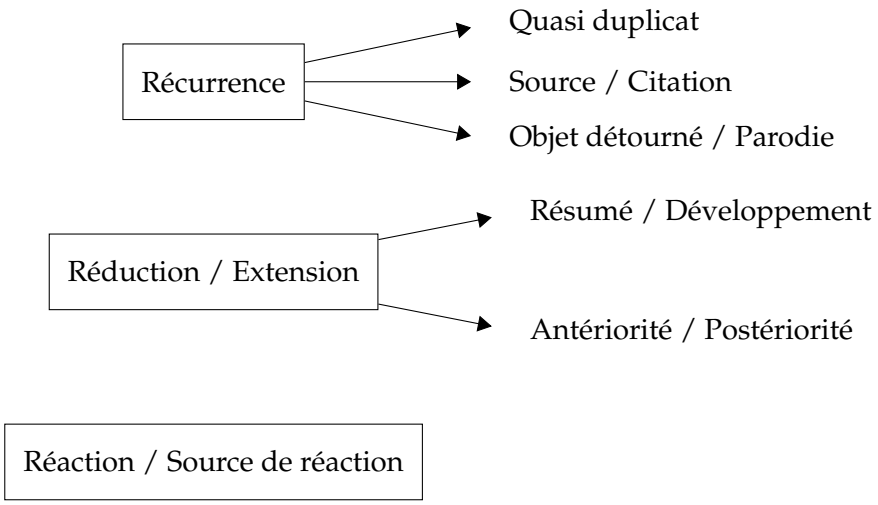


FIGURE 7.1 – Typologie des liens entre informations.

ce/citation et la parodie. Le type « parodie » a été abandonné par manque de documents relevant de ce mode au sein du corpus. En effet, si des émissions de type « Le Petit Journal » ont bien été récupérées, elles étaient la plupart du temps incomplètes, seuls des extraits étant disponibles. Le type « source/citation » a également été abandonné car la mention explicite des sources utilisées dans la presse en ligne est très rare. On dispose au mieux d'une référence au média dont l'information parvient (AFP, concurrent, ...) mais presque jamais de lien HTML vers la source en question. Des travaux de détection d'article originel peuvent être utilisées, mais sont généralement complexes à mettre en œuvre (Allan et al., 2000a). Nous pensons néanmoins que ces deux types sont justifiés, et peuvent présenter un intérêt pour le grand public et les journalistes.

7.1.3 Exemples extraits du corpus

Nous exposons ici deux exemples extraits du corpus. Trois documents sont présentés. Le premier est un article du Figaro daté du 27 février 2015 et rapportant une allocution de Monsieur Manuel Valls lors d'un meeting électoral. Lors de cette allocution, M. Manuel Valls désigne l'extrême droite comme « l'adversaire principal ». Le deuxième est une partie d'une interview de Monsieur Florian Philippot. Durant cet entretien qui se déroule le 28 février 2015, M. Florian Philippot critique l'allocution de M. Manuel Valls et ses propos à l'encontre de son parti. Le troisième est un article du Point qui reprend par écrit l'interview de M. Florian Philippot. Également daté du 28 février 2015, l'article cite sa source et rapporte les paroles de M. Florian Philippot. La figure 7.2 montre les liens existant entre ces documents, en accord avec la typologie décrite précédemment.

La figure 7.3 reprend trois documents illustrant le dépôt de plainte de la ville de Paris après que la chaîne américaine Fox News ait qualifié certains quartiers parisiens de « no-go zones ». L'article du Point présente l'affaire tandis qu'un tweet résume l'article en reprenant l'en-tête tout en citant sa source. L'émission Le Petit Journal parodie l'information en décrivant le dépôt de plainte comme « une bataille entre Madame Anne Hidalgo, maire de Paris, et le premier amendement de la constitution américaine ».

Les liens duaux ne sont pas représentés sur les figures 7.2 et 7.3 dans un souci de lisibilité.

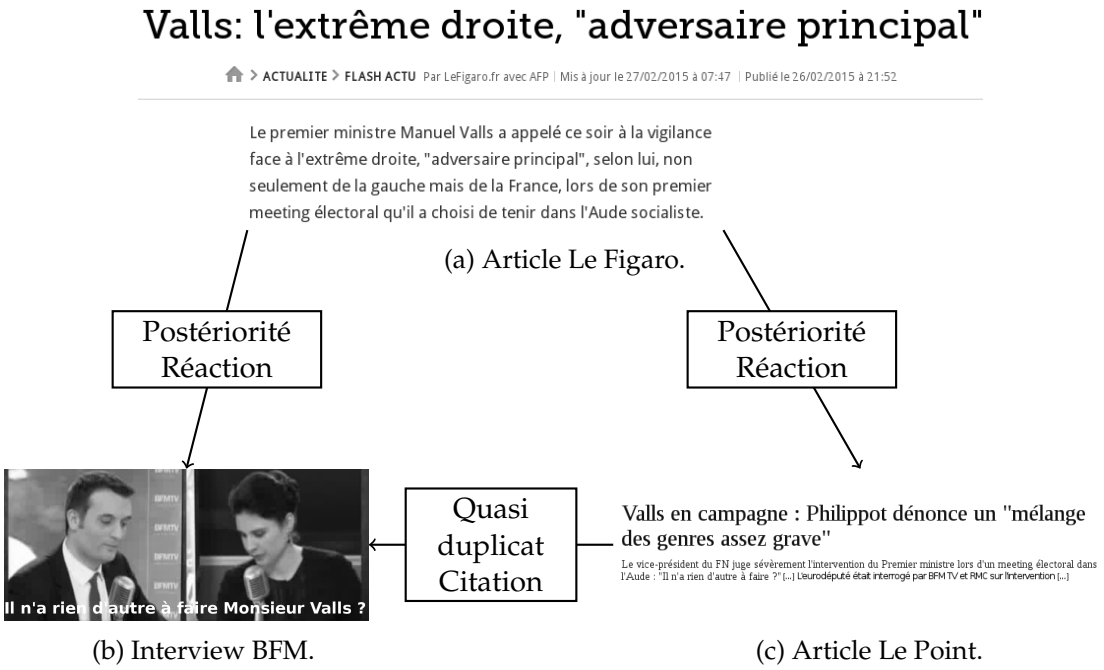


FIGURE 7.2 – Divers liens entre trois informations.

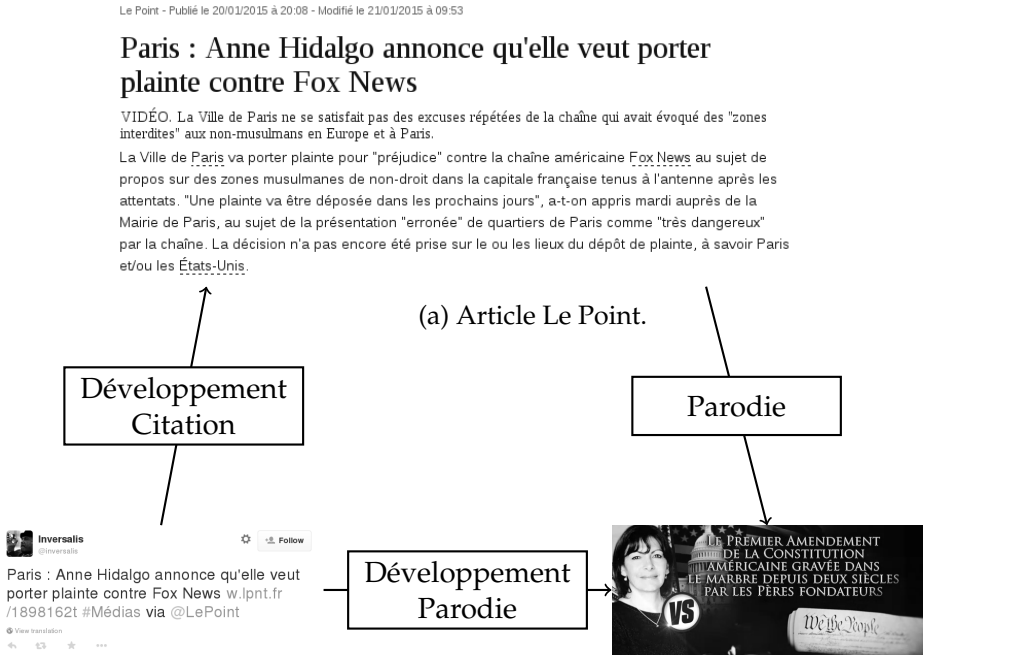


FIGURE 7.3 – Liens de parodie et de développement.

Les liens créés entre ces documents sont indépendants de la modalité de ces derniers. La figure 7.2 montre ainsi un article reprenant les propos tenus par M. Florian Philippot lors d'une interview radiophonique. L'article n'apporte pas davantage d'informations que sa source, et un lien de quasi duplicat est donc créé entre les deux documents bien que leur modalité diffère. On peut néanmoins envisager que certains types de liens soient plus fréquents entre certaines modalités que d'autres.

7.1.4 Ambiguïté du typage

Tout comme il est souvent subjectif de juger de la pertinence d'un lien entre deux documents, il est souvent subjectif de déterminer le type du lien en question. En effet, on peut tout d'abord noter que, pour un seul lien entre un document source et un document cible, plusieurs types pourraient se superposer. Il en est ainsi des réactions, qui sont nécessairement postérieures au document source. On peut néanmoins considérer que la réaction incorpore intrinsèquement cette notion temporelle et est donc un type plus précis que la postériorité, devant être privilégié lorsque cela est possible. Néanmoins, *quid* d'un développement qui inclurait quelques éléments d'informations postérieurs au document source ? Comment typer un résumé qui ne reprendrait que les éléments les plus anciens d'un document source, en supprimant les informations les plus récentes ? Si un typage multiple est possible, sa mise en œuvre créerait des problèmes à la fois d'évaluation (comment décider de la meilleure combinaison de types), et en terme de représentation pour les utilisateurs, qui pourraient être confrontés à de nombreuses combinaisons.

Des expérimentations menées avec des collégiens sur quelques exemples tirés du corpus LIMAH ont montré que si la nature de la relation temporelle était facilement détectée par l'humain, les autres types portaient à débat. En particulier, la notion de la quantité d'informations à supprimer pour obtenir un résumé (et réciproquement à ajouter pour le développement) est difficile à estimer. S'il suffisait d'ôter une unique information pour obtenir un résumé, alors la relation de quasi duplicat ne serait virtuellement jamais instanciée. À l'inverse, imposer un seuil fixe pour déterminer ce qu'est un résumé (*e.g.*, la suppression d'au moins un tiers des informations), pourrait porter à discussion.

Face à ces difficultés, il nous semble illusoire de chercher à assigner « le meilleur type » ou « la meilleure combinaison de types » aux liens, mais qu'il convient plutôt de chercher à éviter les types manifestement incorrects. Ainsi, un lien typé postérieur alors que le document cible rapporte des informations antérieures créerait logiquement une dissonance cognitive chez l'utilisateur. À l'inverse, un lien typé postérieur alors que le document cible rapporte une réaction n'entraînerait pas nécessairement d'inconfort chez l'utilisateur, étant donné que la dimension temporelle est également présente.

7.2 Typage automatique

La typologie proposée plus tôt nous semble pertinente pour le corpus utilisé et le cas d'utilisation envisagé, à savoir l'exploration de collections d'actualités. Appliquée à un jeu de données de plusieurs milliers de documents, il paraît néanmoins indispensable d'automatiser l'étiquetage des liens. Dans cette section, nous décrivons les différentes approches possibles pour cette automatisation et détaillons l'approche retenue.

7.2.1 Approches possibles

Plusieurs approches peuvent être envisager afin de typer automatiquement les liens construits. Une première consiste à utiliser des algorithmes d'apprentissage automatique (*machine learning*) supervisés. Ceux-ci nécessitent un ensemble de liens dont les types ont été annotés par l'humain, ainsi que de caractéristiques (ou *features*) propres aux liens afin d'apprendre un ensemble de règles permettant de passer des caractéristiques à un type. Ainsi, les machines à vecteurs de support (ou *Support Vector Machine, SVMs*) réorganisent l'espace multidimensionnel des caractéristiques afin de trouver un ensemble d'hyperplans permettant de séparer chacun des types (Joachims, 1998). D'autres algorithmes de *machine learning*, de type *K*-NN (Soucy et Mineau, 2001) (dans lequel on assigne à un lien le type du lien le plus similaire dans la collection), ou arbres de décision (Safavian et Landgrebe, 1991) (dans lequel les règles apprises sont explicites) peuvent également être utilisés.

Dans le cas des liens, il est difficile d'extraire des caractéristiques leur étant propres. On peut néanmoins, pour chaque lien, extraire des caractéristiques du document source et du document cible qu'il lie, ainsi que des traits représentant le rapport entre les deux documents. On peut ainsi calculer la similarité lexicale entre les deux documents, le nombre d'entités nommées communes, la présence de marqueurs lexicaux tels que les adverbes de temps ou de lieu, ainsi que des métadonnées telles que la date de publication, l'éditeur, …

Ces approches de *machine learning* sont généralement efficaces pour ce type de tâche. Toutefois, les difficultés d'évaluation décrites plus tôt rendent complexes leurs comparaisons. Notamment, étant donné que plusieurs types sont potentiellement corrects pour chaque lien, une évaluation consistant à vérifier que le type assigné par l'algorithme est l'un des types possibles conduit à privilégier certains types de liens, moins risqués, plutôt que d'autres. Ainsi, les liens temporels risquent fortement d'être sur-représentés.

Pour ces raisons, nous avons choisi une approche par heuristiques, dans laquelle des règles explicites sont établies, permettant d'harmoniser la répartition des différents types de liens. Cette approche et sa mise en œuvre sont décrites dans la section suivante.

7.2.2 Typage à base d'heuristiques

Nous avons retenu une approche à base de règles expertes afin d'obtenir une distribution des types de liens équilibrée. Ainsi, les quelques paramètres à fixer, décrits dans les prochaines lignes, ont été choisis en fonction de la quantité de types qu'ils généraient. Il ne s'agit pas ici de créer des liens, mais de typer des liens déjà détectés. Dans notre cas, la détection en question a été effectuée sur le corpus LIMAH décrit en section 3.2 grâce à l'algorithme des *A*-NN présenté en section 5.3. Pour les 6 812 documents de la collection, ce sont 40 658 liens qui ont été obtenus avec cette méthode, soit en moyenne 6 liens par document. Les types ont ensuite été assignés aux liens dans un ordre spécifique décrit ci-après.

Le premier type de lien attribué est la réaction. Afin de la détecter, deux filtres successifs sont appliqués. D'abord, on vérifie le caractère postérieur de la réaction à une information, en s'assurant grâce aux métadonnées que le document cible du lien ait été publié après le document source. Ensuite, la réaction à proprement parler est détectée à l'aide d'un ensemble de marqueurs lexicaux et syntaxiques tels que les guillemets ou les verbes indiquant une réponse (a réagi, a répondu, …). 2 633 liens de type réaction (5 266

(a) Mots et informations.

(b) Phrases et informations.

FIGURE 7.4 – Corrélation entre le nombre de mots (a) ou de phrases (b) et le nombre d'informations.

si l'on compte les liens inverses) sont ainsi créés, soit environ 13 % des liens.

Le deuxième type assigné correspond aux liens de type temporel. Lors de cette étape, un lien reliant un document source publié au moins un jour avant un document cible se verra attribué le type « antérieur ». De façon réciproque, un lien de type « postérieu » sera assigné pour un document source publié au moins un jour après un document cible. Avec cette méthode, 24 786 liens sont créés, soit 61 % des liens. Cette sur-représentation des liens temporels par rapport aux autres types nous paraît cohérente avec le constat que les liens de temporalité sont les plus largement plébiscités par les journalistes (voir section 1.2.3). Nous choisissons cette approche qui consiste à assigner un type temporel dès que la date de publication diffère pour plusieurs raisons. Tout d'abord, les médias étudiés sont issus de la presse en ligne. Ils sont donc réactifs, et ont un rythme de publication soutenu. Il est donc rare qu'une information parue un jour donné soit reprise sans ajout significatif plus de 24h plus tard. Ensuite, ceci nous permet d'éliminer des types suivants les paires de documents pouvant présenter une évolution d'une information. En effet, le quasi duplicat, le résumé et le développement ne sont envisagés dans notre cas que comme s'appliquant à des documents rapportant des informations similaires, sans évolution, rebondissement ou suivi.

Les développements et résumés sont ensuite détectés grâce à un rapport entre le nombre de mots présents dans chacun des deux documents liés. Si ce rapport est supérieur à 3 pour le développement, ou inférieur à 1/3 pour le résumé, un lien de ce type est créé. Une extraction manuelle des informations d'un échantillon du corpus LIMAH, décrite section 8.2.2, nous a permis d'établir, comme le montre la figure 7.4, que la corrélation entre le nombre de mots et le nombre d'informations est importante. Il semble donc pertinent d'utiliser comme heuristique une différence importante entre nombre de mots afin d'approximer une différence en termes de quantité d'informations. 2 432 liens, pour moitié résumé et pour moitié développement, ont été ainsi créés, soit 6 % des liens.

Enfin, les paires de documents ne répondant à aucun de ces critères se voient assigner le type de quasi duplicat. Ces paires correspondent à des documents de taille similaire et publiés le même jour. Ils représentent environ 20 % des liens, soit 8 174 liens. La répartition en types des 40 658 liens de l'hypergraphe du corpus LIMAH est récapitulé dans la table 7.1.

Type	Quantité	Pourcentage
Antérieur/Postérieur	24 786	61 %
Quasi duplicat	8 174	20 %
Réactions/Sources de réactions	5 266	13 %
Résumé/Développement	2 432	6 %

TABLE 7.1 – Types attribués aux liens de l'hypergraphe LIMAH.

Conclusion

Disposer d'hypergraphes explorables est un prérequis pour offrir la possibilité d'explorer efficacement une collection. Néanmoins, proposer ne serait-ce qu'une dizaine de liens, tous pertinents, à un utilisateur, peut mener à une désorientation. Expliciter la sémantique des liens créés permet de limiter cette désorientation, et autorise l'utilisateur à donner un sens, au propre comme au figuré, à sa navigation. Dans nos travaux, ce typage explicite est exposé par le biais d'une interface graphique, présentée dans le chapitre suivant, au sein de laquelle les liens typés sont directement visibles par l'utilisateur.

Les heuristiques sélectionnées afin de typer automatiquement les liens de l'hypergraphe peuvent sembler grossières. Elles ont néanmoins été choisies dans deux buts : limiter le risque d'erreur en utilisant des paramètres aux valeurs élevées (*e.g.*, le rapport de 1/3 pour la relation de résumé/développement), et assurer une bonne homogénéité des types assignés La pertinence de l'étiquetage est quant à elle complexe à évaluer, étant donné la subjectivité du typage. Face à cette difficulté, il convient de tester cette pertinence de façon extrinsèque, par le biais d'études utilisateurs, ce que nous proposons également dans le chapitre suivant.

Chapitre 8

Validation extrinsèque en situation professionnelle

Introduction

L'hypergraphe typé construit au travers des étapes décrites dans les chapitres 5 et 7 répond à des exigences d'explorabilité et d'aide à la navigation éclairée. Ces exigences ont été évaluées de façon intrinsèque, notamment par une étude de la topologie et des caractéristiques de l'hypergraphe créé (nombre moyen de liens, diamètre, proportion des différents types). Dans l'objectif d'une amélioration de la capacité à explorer l'actualité, et dans l'optique de l'utilisation de telles constructions par des professionnels, il semble primordial de valider ces résultats par une étude extrinsèque, lors de laquelle des utilisateurs sont invités à explorer l'hypergraphe typé.

Dans ce chapitre, nous rapportons les résultats d'une telle étude, menée en coordination avec le CRPCC, et réalisée avec la participation d'étudiants journalistes et de professionnels de la rédaction de Ouest-France. Nous commençons par une description technique et fonctionnelle du système construit ainsi que des interfaces utilisateurs. Nous continuons en détaillant les deux populations étudiées ainsi que le protocole expérimental. Nous concluons par les résultats de cette évaluation extrinsèque, et montrons que l'hypergraphe typé est un outil efficace dans l'aide à la rédaction d'articles de presse.

8.1 Interfaces utilisateur et configurations évaluées

La conception et la réalisation de l'interface utilisateur est une étape déterminante dans la conduite d'une évaluation de fonctionnalités proposées. Une interface peu intuitive, dans laquelle certaines fonctionnalités sont peu visibles ou mal comprises, peut en effet mener à un mauvais ressenti des utilisateurs qui évaluent alors davantage la mauvaise conception de l'interface que les fonctionnalités qu'elle propose. C'est d'autant plus le cas ici étant donné que nous proposons aux utilisateurs une structure avec laquelle la plupart ne sont pas familiers : le graphe. Nous donnons ici un descriptif technique et fonctionnel de l'interface construite pour naviguer au sein de l'hypergraphe LIMAH avant de décrire les 3 versions utilisées lors des tests utilisateurs.

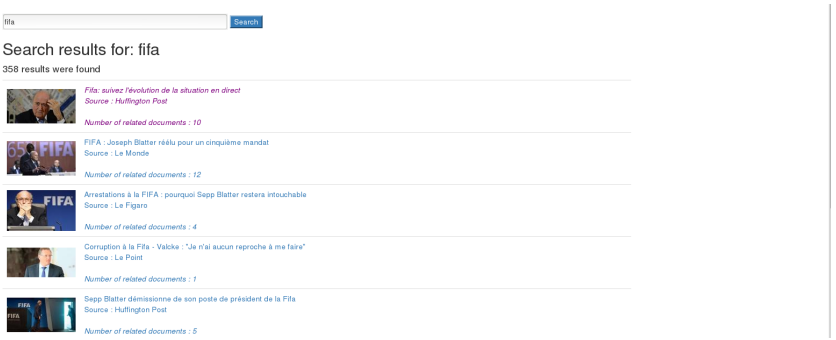


FIGURE 8.1 – Interface « Moteur de recherche ».

FIGURE 8.2 – Interface « Hypergraphe typé ».

8.1.1 Description technique et fonctionnelle

L'interface graphique permettant une visualisation et une navigation au sein de l'hypergraphe a été réalisé par Arnaud Touboulic, ingénieur à l'IRISA au sein du projet LIMAH. Elle se compose de deux parties principales : une interface de recherche par mots-clés basique, qui sert de point d'entrée dans l'hypergraphe, et une interface de navigation permettant de passer d'un document à l'autre sans revenir à une recherche par mots-clés. En saisissant un ou plusieurs mots-clés, l'utilisateur se voit proposer une liste de documents répondant à sa requête, sous une forme typique des moteurs de recherche et présentée en figure 8.1. En sélectionnant l'un des documents qui lui sont proposés, l'utilisateur arrive sur une page présentant le document dans sa totalité, ainsi que divers métadonnées et moyens de navigation.

Cette second interface, illustrée par la figure 8.2, est composée de 5 parties. Au centre, dans la partie basse, se trouve le contenu principal du document. La mise en page originelle des documents web est conservée au mieux, avec l'utilisation des balises de titre, la préservation des liens externes, ou encore la conservation des passages en gras ou en italique. Les documents vidéo sont présentés par le biais d'un lecteur multimédia et d'un texte présentant une transcription automatique du document. Les documents radios bénéficient du même traitement que les documents vidéo.

La partie centrale supérieure représente la vue hypergraphe. Celle-ci permet de na-

viguer entre les différents documents, chacun d'entre eux étant représenté par une icône correspondant au média émetteur. Le document courant, actuellement consulté par l'utilisateur, est centré et clignotant. Les autres documents visibles correspondent à ceux qui sont directement liés au document courant. En d'autres termes, il s'agit des voisins du document courant dans l'hypergraphe tels que détectés par l'algotihme des A-NN. En effectuant un simple clic sur un de ces documents liés, l'interface sera mise à jour afin d'afficher ce nouveau document et ses métadonnées, sans modifier la vue de l'hypergraphe. En cas de double clic sur un document lié, l'interface est mise à jour de façon identique au simple clic, mais actualise également la vue de l'hypergraphe, le nouveau document devenant alors le document courant, ses propres voisins se substituant aux voisins du document précédent. Il reste possible de revenir au document précédent, soit en le trouvant parmi les voisins du nouveau document courant et en double-cliquant dessus, soit en utilisant la touche « précédent » du navigateur web.

Les documents liés sont ordonnés selon plusieurs axes. D'abord, sur l'axe horizontal, les liens de types antérieur et postérieur sont représentés comme une ligne temporelle. Les documents publiés à une même date dans le passé (resp. futur) sont regroupés au sein d'une même boîte, permettant de mieux appréhender l'évolution d'une information et ses moments forts. Au-dessus de chacune de ces boîtes se situe la date de publication des documents. Sous l'icône du document courant, à la verticale, se situent les quasi duplicats. Ils sont également regroupés au sein d'une seule et même boîte. Au-dessus de l'icône du document courant se situent les sources de réaction (diagonale supérieure gauche) et les réactions (diagonale supérieure droite). Lorsque plusieurs réactions ou sources de réactions sont disponibles, elles sont regroupées au sein d'une seule boîte par type, quelque soit leur date de publication. Enfin, en-dessous de l'icône du document, sur les diagonales inférieures gauche et droite, se situent les résumés et développements.

Sur la partie gauche de l'interface, des métadonnées sont présentées. On y trouve notamment la date de publication, l'auteur, le type de document, et un lien vers le document original (sur la page de l'éditeur). On peut également y trouver une liste des mots-clés ordonnés selon leur fréquence dans le document, ainsi qu'une liste des entités nommées (lieux, personnes) apparaissant dans le document. Lors d'un survol des articles liés, les mots-clés partagés par le document survolé et le document courant (*i.e.*, celui affiché dans la partie centrale) sont surlignés.

La partie droite de l'interface est composée de deux parties. La partie supérieure correspond à un filtre en fonction des éditeurs. Ainsi, on peut masquer sur la vue de l'hypergraphe les articles d'un éditeur en particulier (*e.g.*, Le Monde, Le Figaro, ...). Ces articles ne disparaissent pas totalement, mais sont grisés. Sur la partie inférieure droite, des suggestions sont proposées. Celles-ci correspondent aux document représentés sur l'hypergraphe, et sont simplement un autre moyen de représenter cette information. Ces différentes suggestions sont organisées par types (quasi duplicat, réaction, ...) en listes déroulantes.

Une fonctionnalité « déjà lu » a également été implémentée, et se matérialise par l'apparition d'une vignette sur le coin inférieur droit des icônes correspondant à des documents déjà visités par l'utilisateur. Une fonctionnalité de type « favoris » est également présente. Celle-ci permet à l'utilisateur d'enregistrer les documents qui l'intéressent, afin de les retrouver dans une interface dédiée de type « panier ». Cette interface a pour but, dans une version ultérieure, de proposer différents traitements supplémentaires, tels qu'un résumé automatique des documents qui y sont regroupés. Il est par ailleurs possible de surligner des morceaux de textes. Ce surlignage est spécifique à chaque utilisa-

FIGURE 8.3 – Interface 1 : basique.

teur, conservé lors de ses différentes sessions, et peut lui permettre de mettre en avant les passages qu'il estime les plus intéressants.

L'ensemble des documents est stocké au sein d'une base de données NoSQL MongoDB, qui offre directement une fonction de recherche. L'application web est construite grâce à AngularJS et est soutenue par des services REST via Spring et mis en page grâce à Bootstrap. Le *framework* Jhipster a été choisi pour ordonnancer le tout. Afin de représenter l'hypergraphe, D3.js a été utilisé.

8.1.2 Configurations évaluées

Lors de notre étude, nous souhaitons vérifier deux hypothèses. Tout d'abord, que la suggestion de liens, représentés sous la forme d'un hypergraphe tel que décrit précédemment, est utile à l'exploration d'une thématique. Ensuite que le typage des liens apporte une information utile à l'utilisateur, lui permettant d'explorer une collection plus efficacement. Afin de vérifier ces hypothèses, nous utilisons trois versions distinctes de l'interface. La version complète, présentée plus tôt et désormais dénommée « version 3 », sera comparée à deux autres versions, chacune amputée de fonctionnalités.

La version 1 de l'interface, la plus basique, est illustrée par la figure 8.3. Dans cette version, la vue hypergraphe ainsi que toute suggestion de lien est supprimée. L'utilisateur n'a alors plus accès qu'à une visualisation des documents ainsi qu'au moteur de recherche déjà présenté (voir figure 8.1). Les options de filtres et les suggestions, désormais sans objet, sont également masquées. Ce système correspond donc à un moteur de recherche classique, la seule différence étant l'uniformisation de l'esthétique des documents web, ainsi que l'ajout de métadonnées comme les mots clés et les entités nommées. Sa comparaison avec les deux autres versions nous permettra de vérifier l'intérêt d'une visualisation de type hypergraphe pour le parcours d'une collection. Cet intérêt n'est pas évident a priori, les utilisateurs étant très habitués au modèle de moteur de recherche dont ils maîtrisent les codes, mais n'étant, pour la plupart d'entre eux, pas familiers avec la notion de graphe.

La version 2 de l'interface, intermédiaire entre la version 1 « moteur de recherche » et la version 3 complète, propose une vue de l'hypergraphe sans toutefois typer les liens.

FIGURE 8.4 – Interface 2 : hypergraphe non-typé.

La seule relation conservée est une relation temporelle. Ainsi, les réactions et sources de réactions sont ajoutées à la ligne temporelle, et les résumés, développements, et quasi duplicats sont tous réunis sous une seule catégorie pouvant être interprétée comme « les articles publiés le même jour ». La comparaison de cette interface à la version complète nous permettra de vérifier que le typage est une aide pertinente pour l'exploration, notamment en comparaison à une organisation purement temporelle de la collection.

8.2 Populations étudiées et protocole expérimental

Après avoir mené une étude sur les besoins des professionnels (voir en section 1.2.3), il nous a semblé indispensable de fonder l'évaluation de nos systèmes sur une population similaire. Deux populations distinctes ont été étudiées : des étudiants en fin d'études de journalisme, et des professionnels de l'information. Après avoir décrit les caractéristiques de ces deux populations, nous présentons le protocole expérimental utilisé dans cette étude.

8.2.1 Populations étudiées

La première population ayant participé à l'étude utilisateurs est composée de 25 étudiants. Parmi eux, 18 étaient inscrits dans un cursus d'Information-Communication à l'Université de Rennes 2, dont 8 en Licence 3, 8 en Master 1 et 2 en Master 2. Les 7 autres participants étaient inscrits en filière journalisme à l'Institut d'Études Politiques de Rennes, 3 d'entre eux étant en Master 1 et les 4 autres en Master 2. Parmi les étudiants, on compte 12 hommes et 13 femmes, et leur âge moyen s'établit à 24 ans. Les deux parcours dont ces étudiants sont issus mènent à des emplois dans lesquels la manipulation de l'information est quotidienne. Disposant d'une expérience académique ou professionnelle de 3,57 ans en moyenne, ils sont des participants très pertinents à notre étude. En effet, tout en ayant conscience des enjeux et besoins associés à leurs futures professions, ils sont encore dans une phase d'apprentissage qui les rend ouverts à tout nouvel outil pouvant les aider dans leurs futures missions.

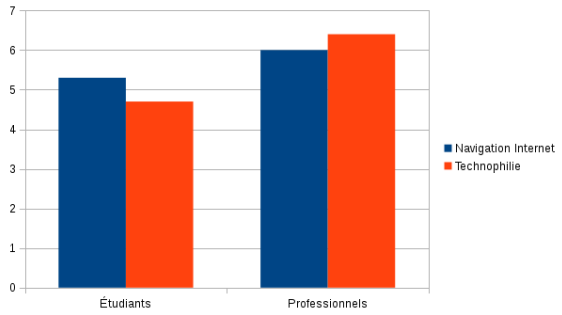


FIGURE 8.5 – Caractéristiques des populations étudiées.

Une étape de pré-questionnaire nous a permis d'évaluer leur appétence pour les nouvelles technologies (désignée ensuite par le terme « technophilie »), ainsi que leurs compétences en navigation Internet. Ces caractéristiques ont été évaluées grâce à une échelle de Likert allant de 1 à 7, 1 signifiant un rejet des nouvelles technologies et 7 une passion. Sur cette échelle, les étudiants obtiennent des scores élevés, avec une moyenne de 5,3 en ce qui concerne la capacité à naviguer efficacement, et à une moyenne de 4,7 en technophilie.

La seconde population étudiée est celle de professionnels de l'information, recrutés au sein de la rédaction de Ouest-France. 6 d'entre eux se sont soumis aux tests, dont 4 hommes et 2 femmes, d'un âge moyen de 37 ans. Leur expérience professionnelle est logiquement plus importante que celle de la population étudiante, avec une moyenne de 11 ans. Parmi eux, on trouve 5 journalistes, dont un *free-lance*, ainsi qu'une documentaliste. Ces professionnels décrivent des compétences perçues en navigation et en technophilie supérieures à celles des étudiants, avec une moyenne 6 pour la navigation et 6,4 pour la technophilie.

Les caractéristiques de technophilie et d'aisance à la navigation internet rapportées par chacune des deux populations étudiées (étudiants et professionnels) sont illustrées figure 8.5.

8.2.2 Protocole expérimental

Les tests réalisés sont composés de quatre parties. Tout d'abord, un entretien préliminaire est conduit afin d'expliquer aux testeurs le contexte de leur participation. Ils remplissent alors un pré-questionnaire permettant d'évaluer leurs compétences en navigation et leur appétence pour les nouvelles technologies, comme décrit plus tôt. Dans une deuxième étape, les utilisateurs visionnent une courte vidéo leur décrivant les fonctionnalités du système qu'ils sont sur le point de tester. Les différentes parties de l'interface (moteur de recherche, métadonnées, suggestions et graphe le cas échéant) leur sont ainsi présentées. La tâche proposée ensuite aux participants consiste à écrire dans un temps limité une synthèse la plus exhaustive possible sur un sujet imposé, à l'aide de l'une des trois interfaces présentées section 8.1. Il est important de noter que les participants à l'étude ne savent pas quelle version de l'interface ils testent, ni même qu'il existe plusieurs versions de cette interface. Ainsi, un utilisateur de la version 1 pensera probablement tester un moteur de recherche enrichi de nouvelles fonctionnalités, tandis que les utilisateurs des versions 2 et 3 constateront plus facilement l'élément innovant, à savoir la possibilité d'une navigation par hypergraphe. Enfin, un post-questionnaire leur

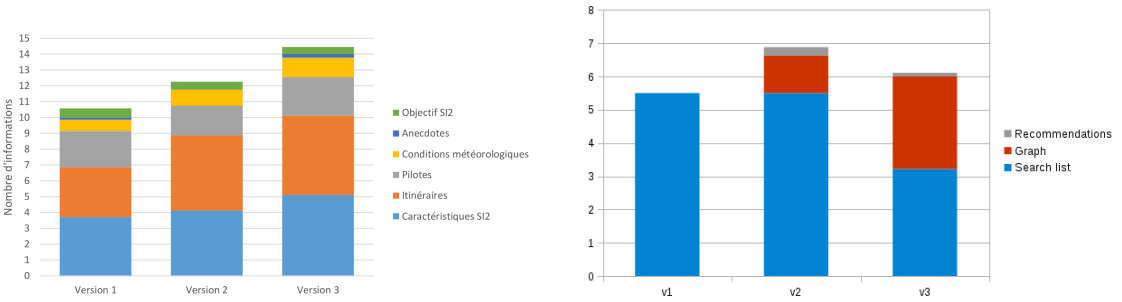
est présenté, et un entretien semi-directif est conduit afin de recueillir leur ressenti sur la version de l'interface qu'ils ont eu l'occasion de tester.

Le sujet imposé aux utilisateurs correspond à une partie du parcours de Solar Impulse 2, un avion solaire piloté par Bertrand Piccard et André Borschberg, et qui a effectué un tour du monde en 2015-2016. Le corpus LIMAH contient plusieurs documents décrivant l'appareil et les pilotes, et se concentrant principalement sur l'une des étapes du tour du monde : un atterrissage imprévu à Nagoya à cause de mauvaises conditions météo. 17 documents au total discutent cette étape. La population étudiante a dû réaliser une synthèse la plus exhaustive possible sur Solar Impulse à l'aide de l'une des trois interfaces. Nous nous sommes assurés lors du pré-questionnaire que la connaissance préalable des participants sur le sujet était limitée, menant à l'éviction d'un unique expérimentateur. Le reste des étudiants a été séparé en trois groupes de 8 à 9 personnes, un groupe pour chaque interface. Ces étudiants avaient pour objectif de remplir la tâche assignée en moins de 20 minutes, un temps suffisamment long pour lire quelques documents en entier, mais suffisamment court pour empêcher une lecture exhaustive des documents disponibles. Les professionnels ont réalisé la même tâche, mais uniquement avec la version 3 de l'interface, la plus complète.

Afin de pouvoir évaluer de façon objective la qualité des résumés obtenus, nous avons procédé avant l'expérimentation à un recueil exhaustif des informations présentes dans le corpus sur Solar Impulse 2. Les 17 documents traitant ce sujet ont donc été manuellement consultés par nos soins, et nous avons pu en extraire 68 informations distinctes, plus ou moins précises (nom et âge du pilote, taille des ailes de l'avion, nombre de capteurs, parcours, ...). Ces informations peuvent être catégorisées comme suit :

— caractéristiques de l'avion (25 informations) ;
— itinéraires (23 informations) ;
— pilotes (14 informations) ;
— conditions météorologiques (3 informations) ;
— anecdotes (2 informations) ;
— objectif du projet (1 information).

La redondance des informations entre les différents documents a également été enregistrée. Ainsi, on a pu compter 13 informations redondantes (présentes dans 7 documents ou plus), 16 informations moyennement redondantes (présentes dans 4 à 6 articles), et 39 informations peu redondantes (présentes dans 1 à 3 documents). De plus, 7 informations ont été catégorisées comme importantes. L'information décrivant l'objectif du projet (la promotion des énergies renouvelables) est ainsi classée comme importante, tout comme les informations représentatives des autres catégories, à l'exception de la catégorie anecdote. Il est à noter que, si l'étude de Cagé et al. (2016) laisse à penser qu'une large partie de la presse en ligne ne publie aucune information originale et se contente de copier un ou quelques articles fondateurs, nous constatons que, pour la très grande majorité des documents, une information unique à la collection est présente. En d'autres termes, la quasi totalité des documents présents dans notre collection rapporte au moins une information originale qui n'est présente dans aucun autre document. Ceci s'est vérifié pour les deux autres thématiques que nous avons explorées (mais qui n'ont pas été utilisées lors de cette étude), à savoir le rachat d'Areva par EDF et la signature du USA Freedom Act limitant la collecte de données de la NSA. Cette constatation ne contredit toutefois pas nécessairement l'étude de Cagé et al. (2016), qui évalue la redondance des chaînes lexicales plutôt que la redondance des informations.



(a) Informations extraites.

(b) Méthode d'accès au document.

FIGURE 8.6 – Résultats en fonction de l'interface utilisée (étudiants).

8.3 Résultats

La constitution de ce protocole expérimental et le contrôle des populations étudiées, en partenariat avec le CRPCC, permet d'envisager des résultats scientifiquement fondés. Nous décrivons ici les résultats de l'étude utilisateur, d'un point de vue quantitatif (nombre et qualité des informations extraites) et qualitatif (ressenti des utilisateurs suite à l'expérimentation).

8.3.1 Évaluation

Tout d'abord, on constate une amélioration significative en termes de nombre d'informations extraites par les utilisateurs de la version 2 par rapport à la version 1, mais également de la version 3 à la version 3. Ainsi, la figure 8.6a montre un nombre moyen d'informations extraites allant de $10,57$ pour la version 1 à $14,44$ pour la version 3. Ces résultats démontrent l'intérêt d'une interface permettant de naviguer de document en document. La figure 8.6b nous permet de constater qu'une structure d'hypergraphe non typé (version 2) a permis aux utilisateurs de parcourir davantage de documents que la version 1. Ce plus grand nombre de documents visités peut expliquer en partie l'amélioration en terme de quantités d'informations extraites entre ces deux versions. Néanmoins, on constate également que les utilisateurs de l'hypergraphe typé (version 3) ont visité moins de documents, tout en extrayant davantage de connaissances que les utilisateurs de l'hypergraphe non typé (version 2).

On remarque que l'ajout de types à l'hypergraphe encourage très largement son utilisation au travers de l'interface. Ainsi, les utilisateurs de la version 2 ont davantage utilisé le moteur de recherche que la navigation au sein de l'hypergraphe, tandis que les utilisateurs de la version 3 se sont servi autant de l'interface d'hypergraphe que du moteur de recherche. Les suggestions, présentées sur le côté droit, ne sont que très peu utilisées sur les deux versions qui la proposent (versions 2 et 3). Ceci peut être expliqué en partie par sa disposition excentrée.

La figure 8.7 montre, pour chaque version du système, la rareté des informations extraites par les utilisateurs. Dans cette évaluation, la version 3 est supérieure aux deux autres versions sur toutes les catégories de rareté des informations. La version 2 est également supérieure à la version 1 dans toutes les catégories à l'exception des informations peu importantes.

La figure 8.8 montre que les professionnels ont obtenu des résultats similaires à ceux

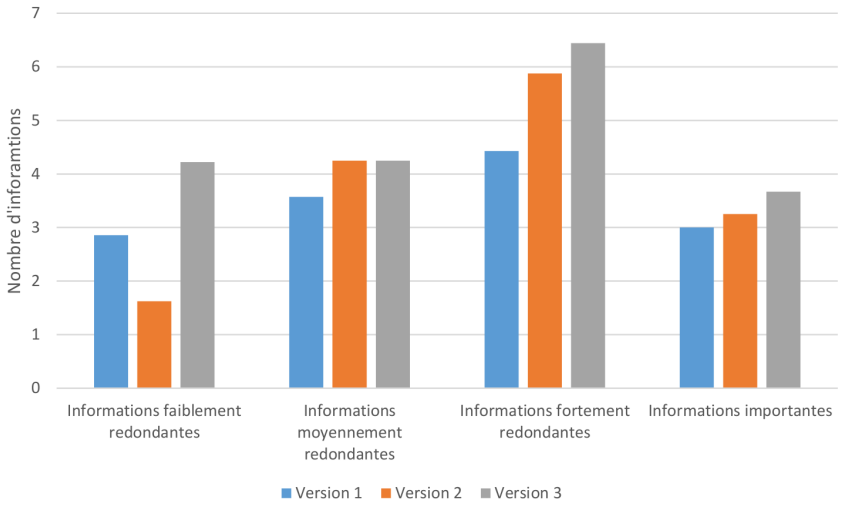


FIGURE 8.7 – Rareté des informations récupérées en fonction de l'interface utilisée (étudiants).

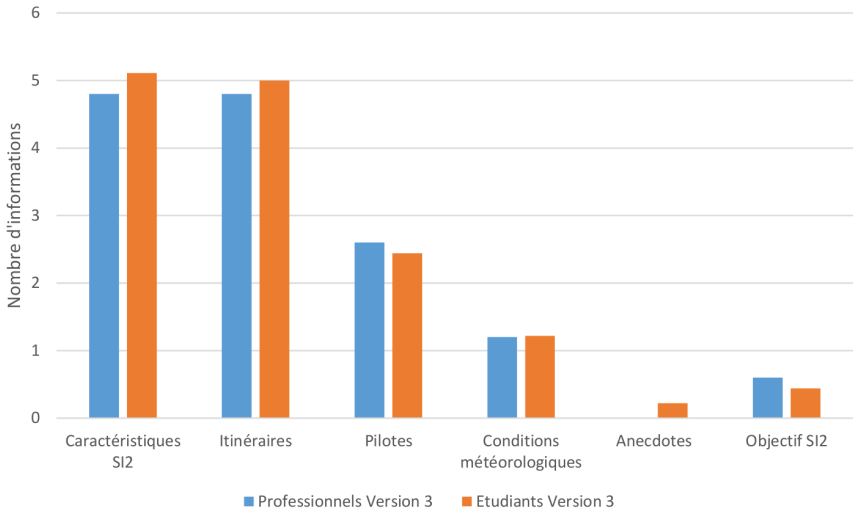


FIGURE 8.8 – Comparaison des informations extraites entre professionnels et étudiants.

des étudiants testés sur la version 3. Il est à noter que les anecdotes n'ont pas été reprises dans les résumés des journalistes. Ceci peut bien entendu être dû au fait que ces informations n'ont pas été trouvées à cause de leur rareté, ou bien au fait qu'elles ont été jugées trop peu importantes pour être notées. Néanmoins, les journalistes professionnels ont en moyenne extrait 4, 2 informations importantes contre 3, 67 pour les étudiants.

Enfin, la figure 8.9 montre que les professionnels ont consulté davantage de documents que les étudiants, et qu'ils ont davantage privilégié l'interface hypergraphe. Ce grand nombre de visites s'explique en partie par le fait que de nombreux journalistes ont cherché à avoir une vue d'ensemble des documents disponibles sur le sujet avant d'en sélectionner un petit nombre à lire de façon plus complète. Les recommandations n'ont quant à elles pas du tout été utilisées par les professionnels.

En résumé, la version 3, comparée aux deux autres, permet d'extraire davantage d'informations, notamment des informations plus rares ou plus importantes. L'interface de

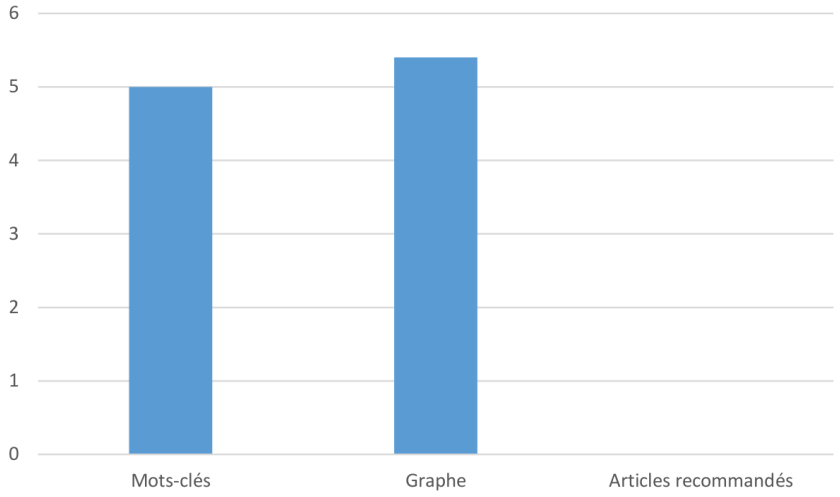


FIGURE 8.9 – Méthode d'accès aux documents (professionnels).

navigation au sein de l'hypergraphe a été assez largement utilisée, mais l'approche par mots-clés est restée une étape importante des recherches effectuées. Ces deux facettes de l'interface proposée semblent donc complémentaires. Ces résultats ont fait l'objet de publications à l'international (Bois et al., 2017b,a).

8.3.2 Ressenti des utilisateurs

Le ressenti des utilisateurs a pu être évalué de deux manières. Dans un premier temps, un post-questionnaire jugeant l'acceptabilité de l'interface testée leur a été remis. Dans un second temps, des entretiens semi-directifs nous ont permis de recueillir leurs remarques et suggestions.

La figure 8.10 montre les réponses des étudiants au post-questionnaire. Il leur a été demandé d'évaluer sur une échelle de Likert allant de 1 à 7 les éléments suivants :

— attentes de performance (*e.g.* « Je trouve que LIMAH est utile pour mon travail. ») ;
— attentes d'effort (*e.g.* « Apprendre à utiliser LIMAH est facile pour moi. ») ;
— motivation hédonique (*e.g.* « Utiliser LIMAH est amusant. ») ;
— intention comportementale (*e.g.* « J'ai l'intention d'utiliser LIMAH dans le futur. »).

Tout d'abord, il est à noter que l'ensemble des scores recueillis pour les trois versions sont élevés. Les attentes de performances sont similaires entre les 3 versions, avec une valeur légèrement moins élevée pour la version 2. On pourrait supposer que l'ajout de types aux liens créés rassure les utilisateurs sur l'efficacité du système, mais les différences observées sont trop faibles pour valider cette hypothèse. Logiquement, la version 1 de type moteur de recherche, familière aux utilisateurs, leur paraît légèrement plus simple à utiliser que les deux autres. Les motivations hédoniques indiquent que la version 1 semble légèrement plus agréable aux utilisateurs. Ceci peut être dû à une interface moins chargée mais, une fois encore, les différences entre les versions sont trop faibles pour pouvoir tirer une conclusion. Enfin, l'évaluation des intentions comportementales indiquent que les utilisateurs se déclarent prêts à utiliser le système dans le futur, cela quelque soit la version utilisée, sans différence significative. On peut penser que la présence de métadonnées telles que les mots-clés et les entités nommées, ainsi que la restriction à des documents journalistiques sont jugées suffisamment intéressantes par les utilisateurs de

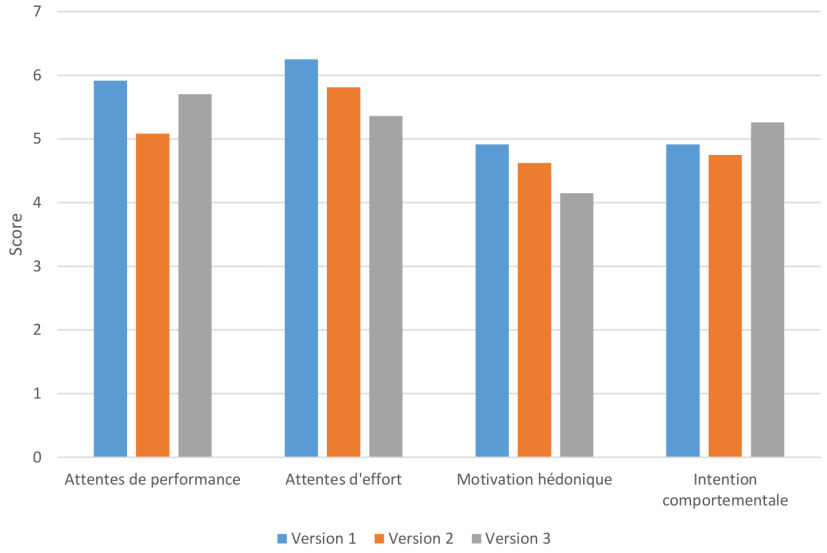


FIGURE 8.10 – Acceptabilité des interfaces (étudiants).

la version 1 pour justifier son utilisation plutôt qu'un moteur de recherche classique. Ces résultats peuvent être jugés comme encourageants étant donné que l'ajout de fonctionnalités innovantes, peu familières aux utilisateurs, ne diminue pas l'attrait du prototype.

Après le remplissage des post-questionnaires, un entretien semi-directif a été conduit avec chacun des participants à l'étude. Après une ouverture de discussion permettant l'évocation spontanée du ressenti de l'utilisateur (« Pouvez-vous me raconter comment s'est passée l'utilisation de l'outil de recherche LIMAH »), il leur a été demandé quels ont été, pour eux, les points forts, points faibles, et axes d'amélioration de l'outil qu'ils ont pu tester.

Concernant la version 1 correspondant à un moteur de recherche classique, les 8 testeurs ont dit avoir apprécié l'utilisabilité de l'interface (7 personnes), l'intérêt des métadonnées (5 personnes), ainsi que le fait que la collection soit spécialisée dans la presse (2 personnes). Divers commentaires positifs sur d'autres fonctionnalités ont également été recueillis (présence des favoris, absence de publicité, efficacité de la recherche par mots-clés). Les points négatifs soulevés pour cette version de l'interface concernent essentiellement des aspects esthétiques (manque d'illustrations), des difficultés d'utilisation (retour au moteur de recherche), ainsi que l'absence de filtres (dates, sources). L'absence d'une source d'information générale de type encyclopédique a également été relevée par un utilisateur, qui aurait souhaité obtenir des informations basiques sur le sujet avant de lire les articles de presse. Parmi les quelques suggestions faites par les utilisateurs et non évoquées plus tôt, on note le besoin d'un court résumé en-dessous des articles dans la liste des résultats du moteur de recherche.

En ce qui concerne la version 2, à savoir l'interface avec un hypergraphe non typé, on note plusieurs retours positifs de la part des 8 testeurs. 5 utilisateurs ont ainsi apprécié l'organisation temporelle de la collection, et estiment que cette fonctionnalité leur a fait gagner du temps. De façon similaire à la version 1, plusieurs utilisateurs ont évalué positivement la présence de métadonnées, qu'il s'agisse de la date et de l'auteur (4 personnes) ou des mots-clés et entités nommées (3 personnes). 3 utilisateurs ont comparé positivement l'interface au moteur de recherche de Google en mettant en avant la fia-

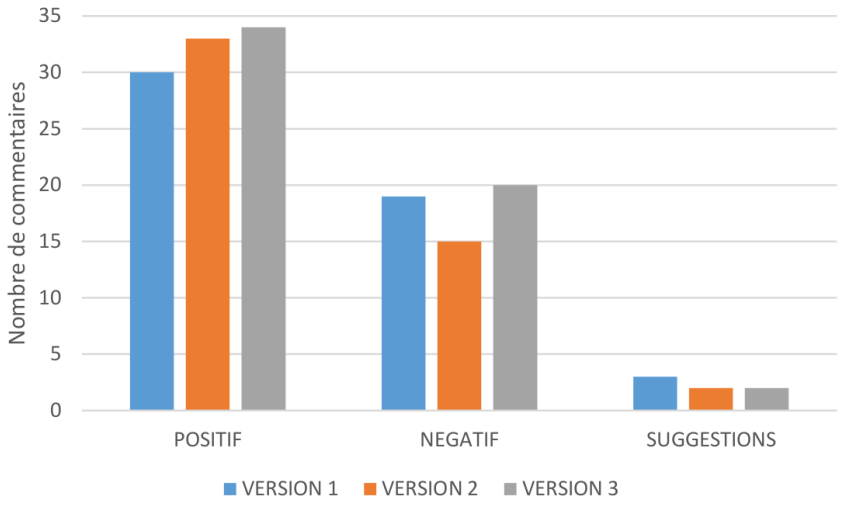


FIGURE 8.11 – Commentaires positifs et négatifs en fonction de l'interface.

bilité et la pertinence des sources disponibles dans la collection. En ce qui concerne les points négatifs, 3 utilisateurs déclarent avoir visité des articles qui n'étaient pas directement lié au sujet. 2 utilisateurs ont exprimé leur déception quant au fait que les articles ne soient liés que par paire (« Je m'attendais à voir plein de liaisons pour pouvoir bien analyser les choses. »). 2 utilisateurs ont également déploré ne pas avoir accès à une source d'information encyclopédique. 1 utilisateur a évoqué sa difficulté à trouver certaines informations précises comme une biographie du pilote. En résumé, pour cette version 2, on constate une bonne acceptabilité de l'interface et une validation de la pertinence d'une organisation temporelle de la collection. Le concept de graphe, avec ses nœuds et ses liens, semble avoir été correctement appréhendé par les expérimentateurs. L'absence de possibilités de recherches précises non limitées à une collection de presse ainsi que l'accès à des connaissances encyclopédiques sur les entités nommées semblent être les principales limites relevées par les utilisateurs.

La version 3 a été testée par 9 personnes. 5 d'entre elles ont estimé que l'outil était pratique et leur avait fait gagner du temps. 4 ont vanté la facilité d'utilisation de l'interface. 3 testeurs ont spécifiquement manifesté leur intérêt pour la représentation graphique des liens entre documents. On notera à ce sujet les deux commentaires suivants : « Le fait que tout soit lié, c'est vraiment pratique et utile pour les recherches et c'est un gain de temps énorme. » et « Ça peut aussi ouvrir la recherche sur des angles auxquels on n'avait pas pensé si on ne connaît pas le sujet. ». Les métadonnées et la fiabilité des sources ont également été positivement évoquées, comme pour les deux autres versions de l'interface (2 personnes). En ce qui concerne les axes d'amélioration, 3 utilisateurs ont déclaré avoir éprouvé des difficultés à revenir sur l'interface de recherche par mots-clés (accessible en cliquant sur la barre de recherche en haut de la page ou via le bouton précédent). 2 utilisateurs ont aussi critiqué la pertinence du moteur de recherche. 2 utilisateurs ont également évoqué le fait qu'ils tombaient sur des documents qui n'étaient pas directement reliés à la thématique en utilisant le moteur de recherche. Comme pour la version 2, quelques utilisateurs ont regretté l'absence de filtres au niveau du moteur de recherche (notamment par date) et un utilisateur aurait souhaité voir davantage de liens sur la représentation hypergraphe. La principale suggestion apportée sur cette interface consiste à donner une indication de la longueur de l'article afin de pouvoir choisir entre des articles longs et

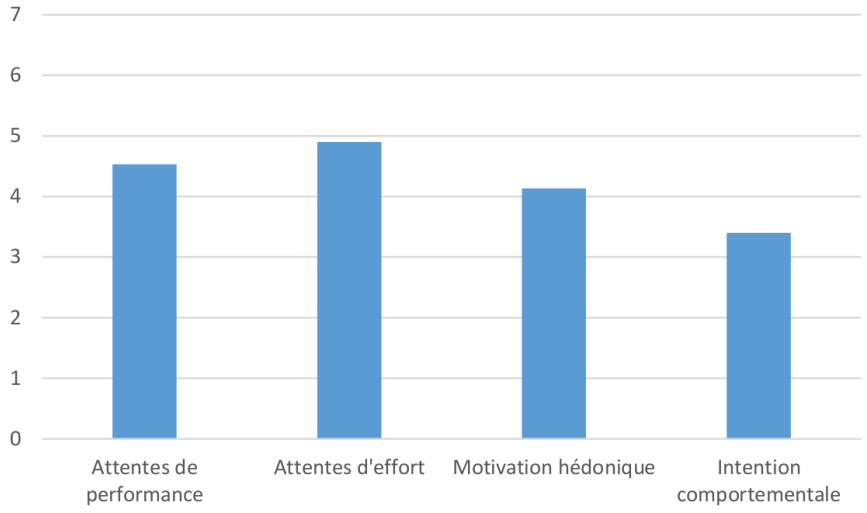


FIGURE 8.12 – Acceptabilité des interfaces (professionnels).

complets ou courts et partiels. En résumé, la version 3 semble avoir été bien acceptée par les utilisateurs. On remarque l'absence totale de commentaires, positifs ou négatifs, sur la pertinence du typage. On peut en conclure que les types choisis sont globalement bien acceptés (ils ne soulèvent pas de questionnement), couvrent les différents aspects de cette thématique (aucun nouveau type n'a été proposé), et que leur assignation est suffisamment pertinente (aucune éventuelle erreur de typage n'a été relevée par les testeurs). Il est toutefois à noter que le type réaction n'est pas instancié sur les documents liés à la thématique de Solar Impulse 2.

La figure 8.11 récapitule le nombre de commentaires positifs ou négatifs des trois versions. On note que la version 3 est la version ayant recueilli le plus de commentaires, positifs comme négatifs, et la version 2 le moins de commentaires négatifs. Le plus grand nombre de commentaires sur la version 3 peut s'expliquer par la présence d'un testeur supplémentaire sur cette interface.

Les professionnels, au nombre de 5, n'ont utilisé que la version 3 de l'interface. Leurs réponses au post-questionnaire sont similaires à celles des étudiants, à l'exception de l'intention comportementale, plus basse, comme montré figure 8.12. Ceci peut s'expliquer par l'existence d'outils davantage ancrés dans les habitudes des professionnels par rapport à celles des étudiants. En termes de retours lors des entretiens semi-directifs, on retrouve de nombreux commentaires, positifs comme négatifs, soulevés par les étudiants. On peut noter que 2 utilisateurs ont manifesté leur intérêt pour la représentation graphique (« Je suis impressionné par l'outil et le fait de voir visuellement les liens. »), et la présence de sources variées (« Ce qui est intéressant avec cet outil, c'est que l'on n'a pas que nos infos à nous (Ouest-France), mais aussi de pouvoir rebondir sur celles d'autres journaux. »). Un utilisateur a également apprécié la présence des recommandations sous forme de listes en déclarant « Nous, avec le titre, on peut rapidement identifier si l'article est pertinent ou pas. ». Cette remarque positive souligne parallèlement la faible visibilité des titres dans l'interface hypergraphe, qui ne sont visibles qu'au survol. 3 utilisateurs ont regretté le fait que tous les articles traitant du sujet ne soient pas réunis en même temps sur le graphe. Cette non exhaustivité est néanmoins prévue par notre approche afin de permettre une exploration plus aisée, lors de laquelle un utilisateur n'est jamais confronté à plusieurs dizaines de documents liés, mais seulement à un faible nombre. 2

testeurs ont également soulevé le fait que la date de publication soit listée dans les métadonnées, sur le côté gauche, alors que cette information primordiale est généralement présente à proximité immédiate du titre de l'article. Plusieurs autres points négatifs repérés par les étudiants ont également été mentionnés par au plus un utilisateur (retour au moteur de recherche, absence de filtre par date, documents non pertinents, absence de contenu encyclopédique).

Conclusion

L'étude réalisée a permis de confirmer la pertinence d'une structuration en hypergraphe, ainsi que celle du typage des liens. Ces résultats encourageants, qui portent à la fois sur la quantité d'informations récupérées grâce à nos systèmes, mais aussi sur la bonne acceptabilité des fonctionnalités, motivent la poursuite de recherches dans ce sens. Le protocole d'évaluation, lourd et coûteux en temps, nous a non seulement fourni ces résultats objectifs, mais a également été l'occasion pour nous d'avoir un aperçu rare de la perception que les professionnels peuvent avoir de tels outils.

L'application de notre approche sur des archives d'actualités telles que celles possédées par les organes de presse semble réalisable, et souhaitable. Son utilité pour le grand public, si elle se justifie aisément d'un point de vue théorique par les caractéristiques topologiques de l'hypergraphe, reste à démontrer aux travers d'études utilisateurs mettant en jeu une exploration nécessitant une dérive thématique, plutôt qu'une synthèse ciblée comme celle qui a été évaluée ici.

Conclusion générale

Nous avons apporté dans cette thèse des contributions à l'état de l'art au travers de deux thématiques principales : l'hyperliage vidéo, et notamment l'amélioration la diversité des systèmes qui le mettent en œuvre, et l'exploration d'actualités.

La première de ces thématiques correspond à un domaine relativement récent, issu de la recherche d'information multimédia, et visant à créer automatiquement des liens entre des segments vidéos arbitraires au sein de grandes collections. Nos participations aux campagnes d'évaluation TRECVid, qui s'intéressent à la problématique de l'hyperliage, nous ont permis de développer des systèmes performants non seulement en terme de pertinence des liens qu'ils proposent, mais également en terme de diversité des liens offerts et de maîtrise de celle-ci. La diversité nous semble être un facteur essentiel et trop souvent ignoré de l'intérêt de tels systèmes. Par la proposition de mesures automatiques permettant de repérer les méthodes offrant peu de diversité, et par la description d'un protocole efficace permettant une comparaison de la diversité des systèmes d'hyperliages, nous avons encouragé l'inclusion de la mesure de la diversité dans les campagnes d'évaluation.

L'application de l'hyperliage à l'ensemble d'une collection est le second angle abordé dans cette thèse. Sa mise en œuvre peut être envisagée de multiples façons, et nous avons défendu ici une approche fondée sur l'explorabilité de l'hypergraphe construit. Nous avons proposé une définition de cette notion d'explorabilité ainsi que des moyens de l'évaluer. Constatant l'inadaptation des algorithmes standards à la création d'une structure répondant aux exigences que nous nous sommes fixées, nous avons développé une nouvelle méthode, les plus proches voisins adaptatifs, disposant de meilleures caractéristiques tant sur le plan de la pertinence des liens créés que sur la topologie du graphe résultant. Une structure explorable nous paraît néanmoins insuffisante pour naviguer efficacement dans des collections hétérogènes de grande taille, multimodales et multisources. Afin de mieux rendre compte à l'utilisateur de la nature du lien qui unie une paire de documents, nous avons proposé et instancié une typologie explicitant la relation existant entre deux documents. La pertinence de la structuration en hypergraphe et du typage des liens qui le composent a été confirmée par le biais d'évaluations humaines impliquant des professionnels de l'information. Cette validation prouve qu'alors que les moteurs de recherche sont aujourd'hui profondément ancrés dans les habitudes des utilisateurs, de nouvelles méthodes innovantes comme celle que nous proposons peuvent se révéler plus efficace dans des contextes professionnels. Cette constatation, nous le pensons, est également valable pour le grand public, qui a tout à gagner à disposer d'outils lui permettant de donner du sens à sa consultation de l'actualité.

Les systèmes décrits dans cette thèse pour la création d'hypergraphes explorables

typés n'ont pas fait l'objet d'un paramétrage extensif. Cette propriété leur confère un intérêt scientifique en terme de reproductibilité, et laisse la voie ouverte à plusieurs améliorations. D'une part, la topologie du graphe obtenu par l'application de l'algorithme des *A*-NN peut être mieux contrôlée. Les quelques composantes non connectées au reste de la structure peuvent y être rattachées grâce à des approches fondées sur l'optimisation sous contraintes permettant de minimiser les liens faibles nouvellement créés. De la même façon, la formation de quelques hubs ou au contraire de nœuds presque isolés, rare dans notre approche mais néanmoins existante, peut être résolue par la combinaison de l'hypergraphe *A*-NN avec des approches de type *K*-NN forçant la création d'un nombre minimal et maximal de liens par nœud. D'autre part, les règles simples utilisées pour le typage automatique des liens peuvent être améliorées par le recours aux approches d'apprentissage automatique. Il semblerait également pertinent d'enrichir la typologie proposée d'un nouveau type représentant les liens dits faibles, potentiellement sérendipitueux, et dont la pertinence peut être remise en cause par l'utilisateur. Cette séparation permettrait non seulement de rassurer l'utilisateur, qui peut croire à un système défectueux lors de la proposition de tels liens pouvant pourtant être jugés pertinents par une tierce personne, mais également de donner à l'utilisateur un meilleur contrôle de sa navigation et notamment de sa potentielle dérive thématique.

Dans notre approche, la similarité entre paires de documents est fondée sur une représentation de surface. Nous avons montré que les approches récentes de type Word2Vec moyenné ne permettaient pas l'application de notre algorithme. Des expériences préliminaires sur les représentations neuronales de documents plus récentes, fondées sur des architectures *Long Short-Term Memory* (LSTM) ou *Gated Reccurrent Unit* (GRU) (Kiros et al., 2015), semblent montrer que l'homogénéité des représentations constatée sur le Word2Vec moyenné y est également présente. Il semble donc pertinent de s'interroger sur la manière dont il serait possible d'inclure cette propriété de brusque chute de similarité aux modèles neuronaux.

Bien que l'outil d'exploration d'actualités que nous avons développé repose sur une collection statique, nous avons montré qu'il était aisément adaptable à une collection dynamique. Néanmoins, nous n'avons envisagé qu'un scénario dans lequel la segmentation des documents se déroule en amont du parcours de la collection. L'hyperliage vidéo, tel que défini dans les campagnes TRECVid, s'envisage d'une manière plus dynamique et porte comme objectif la possibilité de générer un ensemble de suggestions à tout instant du visionnage, portant sur un segment vidéo arbitrairement défini par l'utilisateur. Les interfaces mettant en œuvre de telles fonctionnalités restent aujourd'hui à inventer. Les approches fondées sur l'utilisation d'un second écran semblent particulièrement adaptée à ce cas d'usage et constituent sans aucun doute une piste prometteuse.

Enfin, nous nous sommes limités dans ces travaux à l'exploitation d'une collection d'actualités. Si cette dernière est riche d'un point de vue informationnel de par ses caractéristiques multisources et multimodales, il semble indispensable d'y associer des ressources externes permettant une meilleure compréhension des éléments qui y sont mentionnés. On peut notamment penser aux ressources encyclopédiques permettant une description efficace des entités nommées, mais également à l'*open data*, aux textes de lois, aux rapports parlementaires, qui peuvent mettre en lumière ou développer des aspects qui ne seraient qu'abordés brièvement dans un document journalistique.

Bibliographie

Lada Adamic et Eytan Adar. How to Search a Social Network. *Social networks*, 27(3) : 187–203, 2005.

Lada A Adamic. The Small World Web. In *Proceedings of the International Conference on Theory and Practice of Digital Libraries*, pages 443–452. Springer, 1999.

Amr Ahmed, Qirong Ho, Jacob Eisenstein, Eric Xing, Alexander J Smola, et Choon Hui Teo. Unified Analysis of Streaming News. In *Proceedings of the International Conference on World Wide Web*, pages 267–276, 2011.

James Allan, Jaime Carbonell, George Doddington, Jonathan Yamron, Yiming Yang, James Allan Umass, Brian Archibald Cmu, Doug Beeferman Cmu, Adam Berger Cmu, Ralf Brown Cmu, et al. Topic Detection and Tracking Pilot Study Final Report. In *Proceedings of the Workshop on Broadcast News Transcription and Understanding*, 1998.

James Allan, Victor Lavrenko, et Hubert Jin. First Story Detection in TDT is Hard. In *Proceedings of the International Conference on Information and Knowledge Management*, pages 374–381. ACM, 2000a.

James Allan, Victor Lavrenko, Daniella Malin, et Russell Swan. Detections, Bounds, and Timelines : Umass and TDT-3. In *Proceedings of the International Workshop on Topic Detection and Tracking*, pages 167–174. sn, 2000b.

Hunt Allcott et Matthew Gentzkow. Social Media and Fake News in the 2016 Election. Technical report, National Bureau of Economic Research, 2017.

Omar Alonso, Michael Gertz, et Ricardo Baeza-Yates. Clustering and Exploring Search Results Using Timeline Constructions. In *Proceedings of the International Conference on Information and Knowledge Management*, pages 97–106. ACM, 2009.

Susan Athey et Markus Mobius. The Impact of News Aggregators on Internet News Consumption : The Case of Localization. Technical report, 2012.

George Awad, Jonathan Fiscus, Martial Michel, David Joy, Wessel Kraaij, Alan F Smeaton, Georges Quénot, Maria Eskevich, Robin Aly, et Roeland Ordelman. TRECVID 2016 : Evaluating Video Search, Video Event Detection, Localization, and Hyperlinking. In *Proceedings of the TRECVID Workshop*, volume 2016, 2016.

Joel Azzopardi et Christopher Staff. Incremental Clustering of News Reports. volume 5, pages 364–378. Molecular Diversity Preservation International, 2012.

Hrvoje Bacan, Igor S Pandzic, et Darko Gulija. Automated News Item Categorization. In *Proceedings of the Conference of The Japanese Society for Artificial Intelligence*, pages 251–256, 2005.

Lars Backstrom et Jure Leskovec. Supervised random walks : predicting and recommending links in social networks. In *Proceedings of the International Conference on Web Search and Data Mining*, pages 635–644. ACM, 2011.

Eytan Bakshy, Itamar Rosenn, Cameron Marlow, et Lada Adamic. The Role of Social Networks in Information Diffusion. In *Proceedings of the International Conference on World Wide Web*, pages 519–528. ACM, 2012.

Marko Balabanović et Yoav Shoham. Fab : Content-based, Collaborative Recommendation. *Communications of the ACM*, 40(3) :66–72, 1997.

Juan Manuel Barrios, Jose M Saavedra, Felipe Ramirez, et David Contreras. ORAND at TRECVID 2015 : Instance Search and Video Hyperlinking Tasks. In *Proceedings of the TRECVID Workshop*, 2015.

Marcia J Bates. The Design of Browsing and Berrypicking Techniques for the Online Search Interface. *Online review*, 13(5) :407–424, 1989.

Chidansh Bhatt, Nikolaos Pappas, Maryam Habibi, et Andrei Popescu-Belis. Idiap at MediaEval 2013 : Search and Hyperlinking Task. In *Proceedings of the MediaEval Workshop*, 2013.

Giang Binh Tran, Mohammad Alrifai, et Dat Quoc Nguyen. Predicting Relevant News Events for Timeline Summaries. In *Proceedings of the International Conference on World Wide Web*, pages 91–92. ACM, 2013.

David M. Blei, Andrew Y. Ng, et Michael I. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3 :993–1022, 2003. ISSN 1532-4435. URL http://dl.acm.org/citation.cfm?id=944919.944937.

Toine Bogers et Antal Van den Bosch. Comparing and Evaluating Information Retrieval Algorithms for News Recommendation. In *Proceedings of the Conference on Recommender Systems*, pages 141–144. ACM, 2007.

Rémi Bois, Guillaume Gravier, Pascale Sébillot, et Emmanuel Morin. Vers une Typologie de Liens entre Contenus Journalistiques. In *Proceedings of the Conference on Traitement Automatique des Langues Naturelles*, pages 515–521, 2015.

Rémi Bois, Vedran Vukotić, Ronan Sicre, Christian Raymond, Guillaume Gravier, et Pascale Sébillot. IRISA at TRECVid2016 : Crossmodality, Multimodality and Monomodality for Video Hyperlinking. In *Proceedings of the TRECVID Workshop*, 2016.

Rémi Bois, Guillaume Gravier, Éric Jamet, Emmanuel Morin, Maxime Robert, et Pascale Sébillot. Linking Multimedia Content for Efficient News Browsing. In *Proceedings of the International Conference on Multimedia Retrieval*, pages 301–307, 2017a.

Rémi Bois, Guillaume Gravier, Eric Jamet, Emmanuel Morin, Pascale Sébillot, et Maxime Robert. Language-based Construction of Explorable News Graphs for Journalists. In *Proceedings of the International Workshop : Natural Language Processing meets Journalism*, pages 31–36, 2017b.

Rémi Bois, Vedran Vukotić, Anca-Roxana Simon, Ronan Sicre, Christian Raymond, Pascale Sébillot, et Guillaume Gravier. Exploiting Multimodality in Video Hyperlinking to Improve Target Diversity. In *Proceedings of the International Conference on Multimedia Modeling*, pages 185–197. Springer, Cham, 2017c.

Ilaria Bordino, Yelena Mejova, et Mounia Lalmas. Penguins in Sweaters, or Serendipitous Entity Search on User-generated Content. In *Proceedings of the International Conference on Information & Knowledge Management*, pages 109–118. ACM, 2013.

David B Bracewell, Jiajun Yan, Fuji Ren, et Shingo Kuroiwa. Category Classification and Topic Discovery of Japanese and English News Articles. *Electronic Notes in Theoretical Computer Science*, 225 :51–65, 2009.

Paul S Bradley et Usama M Fayyad. Refining Initial Points for K-Means Clustering. In *Proceedings of the International Conference on Machine Learning*, volume 98, pages 91–99, 1998.

Paul Bradshaw. What is Data Journalism. *Ethics for Digital Journalists : Emerging Best Practices*, pages 202–219, 2014.

Martin G Brown, Jonathon T Foote, Gareth JF Jones, Karen Sparck Jones, et Steve J Young. Automatic Content-based Retrieval of Broadcast News. In *Proceedings of the International Conference on Multimedia*, pages 35–43. ACM, 1995.

Peter Brusilovsky. Methods and Techniques of Adaptive Hypermedia. In *Adaptive hypertext and hypermedia*, pages 1–43. Springer, 1998.

Erik P Bucy. Second Generation Net News : Interactivity and Information Accessibility in the Online Environment. *International Journal on Media Management*, 6(1-2) :102–113, 2004.

Julia Cagé, Nicolas Hervé, et Marie-Luce Viaud. The Production of Information in an Online World : Is Copy Right ? *Working Paper*, 2016.

Julia Cagé, Nicolas Hervé, et Marie-Luce Viaud. *The Production of Information in an Online World : Is Copy Right ?* INA, 2017.

Joan Calzada et Ricard Gil. What Do News Aggregators Do ? Evidence from Google News in Spain and Germany. *Working Paper*, 2016.

Jaime Carbonell, Yiming Yang, John Lafferty, Ralf D Brown, Tom Pierce, et Xin Liu. CMU report on TDT-2 : Segmentation, detection and tracking. In *Proceedings of the Workshop on Broadcast News*, pages 117–120, 1999.

Zhiyong Cheng, Xuanchong Li, Jialie Shen, et Alexander G. Hauptmann. CMU-SMU@TRECVID 2015 : Video Hyperlinking. In *Proceedings of the TRECVID Workshop*, 2015.

Pascal R Chesnais, Matthew J Mucklo, et Jonathan A Sheena. The Fishwrap Personalized News System. In *Proceedings of the International Workshop on Community Networking*, pages 275–282. IEEE, 1995.

Rohan Choudhary, Sameep Mehta, Amitabha Bagchi, et Rahul Balakrishnan. Towards Characterization of Actor Evolution and Interactions in News Corpora. In *Proceedings of the European Conference on Information Retrieval*, pages 422–429. Springer, 2008.

Michael Christel, Scott Stevens, et Howard Wactlar. Informedia Digital Video Library. In *Proceedings of the International Conference on Multimedia*, pages 480–481. ACM, 1994.

Charles LA Clarke, Maheedhar Kolla, Gordon V Cormack, Olga Vechtomova, Azin Ashkan, Stefan Büttcher, et Ian MacKinnon. Novelty and Diversity in Information Retrieval Evaluation. In *Proceedings of the International Conference on Research and Development in Information Retrieval*, pages 659–666. ACM, 2008.

Chip Cleary et Ray Bareiss. Practical Methods for Automatically Generating Typed Links. In *Proceedings of the International Conference on Hypertext*, pages 31–41. ACM, 1996.

Mark Coddington. Clarifying Journalism's Quantitative Turn : A Typology for Evaluating Data Journalism, Computational Journalism, and Computer-Assisted Reporting. *Digital Journalism*, 3(3) :331–348, 2015.

Anca-Roxana Şimon, Guillaume Gravier, Pascale Sébillot, et Marie-Francine Moens. IRISA and KUL at MediaEval 2014 : Search and Hyperlinking Task. In *Proceedings of the MediaEval Workshop*, 2014.

Anca-Roxana Şimon, Rémi Bois, Guillaume Gravier, Pascale Sébillot, Emmanuel Morin, et Sien Moens. Hierarchical Topic Models for Language-based Video Hyperlinking. In *Proceedings of the International Workshop on Speech, Language and Audio in Multimedia*, 2015.

Xiang-Ying Dai, Qing-Cai Chen, Xiao-Long Wang, et Jun Xu. Online Topic Detection and Tracking of Financial News Based on Hierarchical Clustering. In *Proceedings of the International Conference on Machine Learning and Cybernetics*, volume 6, pages 3341–3346. IEEE, 2010.

Maximilien Danisch, Jean-Loup Guillaume, et Bénédicte Le Grand. Towards Multi-egocentred Communities : a Node Similarity Approach. *International Journal of Web Based Communities*, 9(3) :299–322, 2013.

Hal Daumé III et Daniel Marcu. Bayesian Query-focused Summarization. In *Proceedings of the International Conference on Computational Linguistics*, pages 305–312. Association for Computational Linguistics, 2006.

Nick Davies. *Flat Earth News*. Random House UK, 2009.

Tom De Nies, Wesley De Neve, Erik Mannens, et Rik Van de Walle. Ghent University-iMinds at MediaEval 2013 : an Unsupervised Named Entity-based Similarity Measure for Search and Hyperlinking. In *Proceedings of the MediaEval Workshop*, 2013.

Ork de Rooij et Marcel Worring. Browsing Video Along Multiple Threads. *IEEE Transactions on Multimedia*, 12(2) :121–130, 2010.

Gianna M Del Corso, Antonio Gulli, et Francesco Romani. Ranking a Stream of News. In *Proceedings of the International Conference on World Wide Web*, pages 97–106. ACM, 2005.

Chris HQ Ding, Xiaofeng He, Hongyuan Zha, Ming Gu, et Horst D Simon. A Min-max Cut Algorithm for Graph Partitioning and Data Clustering. In *Proceedings of the International Conference on Data Mining*, pages 107–114. IEEE, 2001.

Abdigani Diriye, Srdan Zagorac, Suzanne Little, et Stefan Rüger. NewsRoom : An Information-seeking Support System for News Videos. In *Proceedings of the International Conference on Multimedia Information Retrieval*, pages 377–380. ACM, 2010.

Anhai Doan, Raghu Ramakrishnan, et Alon Y Halevy. Crowdsourcing Systems on the World-wide Web. *Communications of the ACM*, 54(4) :86–96, 2011.

Paul Ekman. An Argument for Basic Emotions. *Cognition & emotion*, 6(3-4) :169–200, 1992.

Wander Emediato. L'Argumentation dans le Discours d'Information Médiatique. *Argumentation et Analyse du Discours*, (7), 2011.

Levent Ertöz, Michael Steinbach, et Vipin Kumar. Finding Topics in Collections of Documents : A Shared Nearest Neighbor Approach. In *Clustering and Information Retrieval*, pages 83–103. Springer, 2004.

Olivier Ertzscheid. *Le Lieu, le Lien, le Livre : les Enjeux Cognitifs et Stylistiques de l'Organisation Hypertextuelle*. PhD thesis, Université de Toulouse 2, 2002.

Maria Eskevich, Gareth JF Jones, Shu Chen, Robin Aly, Roeland Ordelman, et Martha Larson. Search and hyperlinking task at MediaEval 2012. *CEUR Workshop Proceedings*, 927, 2012.

Maria Eskevich, Gareth JF Jones, Robin Aly, Roeland JF Ordelman, Shu Chen, Danish Nadeem, Camille Guinaudeau, Guillaume Gravier, Pascale Sébillot, Tom De Nies, et al. Multimedia Information Seeking Through Search and Hyperlinking. In *Proceedings of the International Conference Multimedia Retrieval*, pages 287–294. ACM, 2013.

Maria Eskevich, Robin Aly, David N. Racca, Roeland Ordelman, Shu Chen, et Gareth J. F. Jones. The Search and Hyperlinking Task at MediaEval 2014. In *Proceedings of the MediaEval Workshop*, 2014.

Maria Eskevich, Martha Larson, Robin Aly, Serwah Sabetghadam, Gareth J. F. Jones, Roeland Ordelman, et Benoit Huet. Multimodal Video-to-Video Linking : Turning to the Crowd for Insight and Evaluation. In *Proceedings of the International Conference on Multimedia Modeling*, 2017.

Benoit Favre, Jean-François Bonastre, et Patrice Bellot. Recherche d'information dans un mélange de documents écrits et parlés. In *Proceedings of the Journées d'Etude de la Parole*, pages 403–412, 2004.

Michael R Fellows, Jiong Guo, Christian Komusiewicz, Rolf Niedermeier, et Johannes Uhlmann. Graph-based data clustering with overlaps. *Discrete Optimization*, 8(1) :2–17, 2011.

Fangxiang Feng, Xiaojie Wang, et Ruifan Li. Cross-modal Retrieval with Correspondence Autoencoder. In *Proceedings of the International Conference on Multimedia*, pages 7–16. ACM, 2014.

Natalie Fenton. *New Media, Old News : Journalism and Democracy in the Digital Age*. Sage Publications, 2010.

Katherine Fink et Michael Schudson. The Rise of Contextual Journalism, 1950s–2000s. *Journalism*, 15(1) :3–20, 2014.

Rana Forsati, Mehrdad Mahdavi, Mehrnoush Shamsfard, et Mohammad Reza Meybodi. Efficient Stochastic Algorithms for Document Clustering. *Information Sciences*, 220 : 269–291, 2013.

Allen Foster et Nigel Ford. Serendipity and Information Seeking : an Empirical Study. *Journal of Documentation*, 59(3) :321–340, 2003.

Amel Fraisse et Patrick Paroubek. Toward a Unifying Model for Opinion, Sentiment and Emotion Information Extraction. In *Proceedings of the International Conference on Language Resources and Evaluation*, pages 3881–3886, 2014.

David Frey, Rahul Gupta, Vikas Khandelwal, Victor Lavrenko, Anton Leuski, et James Allan. Monitoring the News : a TDT Demonstration System. In *Proceedings of the International Conference on Human Language Technology Research*, 2001. URL http://aclweb.org/anthology/H01-1053.

Giorgio Gallo, Giustino Longo, Stefano Pallottino, et Sang Nguyen. Directed Hypergraphs and Applications. *Discrete Applied Mathematics*, 42(2-3) :177–201, 1993.

Petra Galuscáková, Martin Krulis, Jakub Lokoc, et Pavel Pecina. CUNI at MediaEval 2014 Search and Hyperlinking Task : visual and Prosodic Features in Hyperlinking. In *Proceedings of the MediaEval Workshop*, 2014.

Fabio Gasparetti. Modeling User Interests from Web Browsing Activities. *Data Mining and Knowledge Discovery*, pages 1–46, 2016.

Jean-Luc Gauvain, Lori Lamel, et Gilles Adda. The LIMSI Broadcast News Transcription System. *Speech Communication*, 37(1-2) :89–108, 2002.

Mouzhi Ge, Carla Delgado-Battenfeld, et Dietmar Jannach. Beyond Accuracy : Evaluating Recommender Systems by Coverage and Serendipity. In *Proceedings of the Conference on Recommender Systems*, pages 257–260. ACM, 2010.

Homero Gil de Zúñiga, Nakwon Jung, et Sebastián Valenzuela. Social Media Use for News and Individuals' Social Capital, Civic Engagement and Political Participation. *Journal of Computer-Mediated Communication*, 17(3) :319–336, 2012.

Goran Glavaš et Jan Šnajder. Event Graphs for Information Retrieval and Multi-document Summarization. *Expert Systems with Applications*, 41(15) :6904–6916, 2014.

Wael H Gomaa et Aly A Fahmy. A Survey of Text Similarity Approaches. *International Journal of Computer Applications*, 68(13), 2013.

Nathaniel Good, J Ben Schafer, Joseph A Konstan, Al Borchers, Badrul Sarwar, Jon Herlocker, John Riedl, et al. Combining Collaborative Filtering with Personal Agents for Better Recommendations. In *Proceedings of the Conference on Innovative Applications of Artificial Intelligence*, pages 439–446, 1999.

Jeffrey Gottfried et Elisa Shearer. News Use across Social Media Platforms 2016. *Pew Research Center*, 26 :3, 2016.

Lucas Graves. The Affordances of Blogging : A Case Study in Culture and Technological Effects. *Journal of Communication Inquiry*, 31(4) :331–346, 2007.

Lucas Graves et Federica Cherubini. The Rise of Fact-checking Sites in Europe, 2016.

Guillaume Gravier, Martin Ragot, Laurent Amsaleg, Rémi Bois, Grégoire Jadi, Eric Jamet, Laura Monceaux, et Pascale Sébillot. Shaping-Up Multimedia Analytics : Needs and Expectations of Media Professionals. In *Proceedings of the International Conference on Multimedia Modeling , Perspectives on Multimedia Analytics*, 2016.

Camille Guinaudeau, Guillaume Gravier, et Pascale Sébillot. IRISA at MediaEval 2012 : Search and Hyperlinking Task. In *Proceedings of the MediaEval Workshop*, 2012a.

Camille Guinaudeau, Guillaume Gravier, et Pascale Sébillot. Enhancing lexical cohesion measure with confidence measures, semantic relations and language model interpolation for multimedia spoken content topic segmentation. *Computer Speech & Language*, 26(2) :90–104, 2012b.

Barrie Gunter. *News and the Net*, volume 9. Routledge, 2003.

Alan Haggerty, Ryen W White, et Joemon M Jose. NewsFlash : Adaptive TV News Delivery on the Web. In *Proceedings of the International Workshop on Adaptive Multimedia Retrieval*, pages 72–86. Springer, 2003.

Kazi Saidul Hasan et Vincent Ng. Conundrums in Unsupervised Keyphrase Extraction : Making Sense of the State-of-the-art. In *Proceedings of the International Conference on Computational Linguistics*, 2010.

Vasileios Hatzivassiloglou, Luis Gravano, et Ankineedu Maganti. An Investigation of Linguistic Features and Clustering Algorithms for Topical Document Clustering. In *Proceedings of the International Conference on Research and Development in Information Retrieval*, pages 224–231. ACM, 2000.

AG Hauptmann, P Scheytt, HD Wactlar, et PE Kennedy. Multi-lingual Informedia : a Demonstration of Speech Recognition and Information Retrieval across Multiple Languages. In *Proceedings of the Workshop on Broadcast News Transcription and Understanding*, 1998.

Alexander Hauptmann. Lessons for the Future from a Decade of Informedia Video Analysis Research. *Image and video retrieval*, pages 595–595, 2005.

Alexander G Hauptmann et Michael J Witbrock. Informedia : News-on-demand Multimedia Information Acquisition and Retrieval. *Intelligent Multimedia Information Retrieval*, pages 215–239, 1997.

Leonhard Hennig et DAI Labor. Topic-based Multi-Document Summarization with Probabilistic Latent Semantic Analysis. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, pages 144–149, 2009.

Jack Herbert et Neil Thurman. Paid Content Strategies for News Websites : An Empirical Study of British Newspapers' Online Business Models. *Journalism practice*, 1(2) :208–226, 2007.

Geoffrey E Hinton et Ruslan R Salakhutdinov. Reducing the Dimensionality of Data with Neural Networks. *Science*, 313(5786) :504–507, 2006.

Avery E Holton et Hsiang Iris Chyi. News and the Overloaded Consumer : Factors Influencing Information Overload among News Consumers. *Cyberpsychology, Behavior, and Social Networking*, 15(11) :619–624, 2012.

Kai Hong, John M Conroy, Benoit Favre, Alex Kulesza, Hui Lin, et Ani Nenkova. A Repository of State of the Art and Competitive Baseline Summaries for Generic News Summarization. In *Proceedings of the International Conference on Language Resources and Evaluation*, pages 1608–1616, 2014.

Leo Iaquinta, Marco De Gemmis, Pasquale Lops, Giovanni Semeraro, Michele Filannino, et Piero Molino. Introducing Serendipity in a Content-Based Recommender System. In *Proceedings of the Conference on Hybrid Intelligent Systems Conference*, pages 168–173. IEEE, 2008.

Ichiro Ide, Hiroshi Mo, Norio Katayama, et Shin'ichi Satoh. Topic Threading for Structuring a Large-scale News Video Archive. In *Proceedings of the International Conference on Image and Video Retrieval*, pages 123–131. Springer, 2004.

Ichiro Ide, Tomoyoshi Kinoshita, Tomokazu Takahashi, Hiroshi Mo, Norio Katayama, Shin'ichi Satoh, et Hiroshi Murase. Efficient Tracking of News Topics Based on Chronological Semantic Structures in a Large-scale News Video Archive. *IEICE Transactions on Information and Systems*, 95(5) :1288–1300, 2012.

Wouter IJntema, Frank Goossen, Flavius Frasincar, et Frederik Hogenboom. Ontology-based News Recommendation. In *Proceedings of the International Conference on Extending Database Technology/International Conference on Database Theory Workshops*, page 16. ACM, 2010.

Piotr Indyk et Rajeev Motwani. Approximate Nearest Neighbors :Ttowards Removing the Curse of Dimensionality. In *Proceedings of the Symposium on Theory of Computing*, pages 604–613. ACM, 1998.

Thorsten Joachims. Text Categorization with Support Vector Machines : Learning with Many Relevant Features. *European Conference on Machine Learning*, pages 137–142, 1998.

Junzo Kamahara, Tomofumi Asakawa, Shinji Shimojo, et Hideo Miyahara. A Community-based Recommendation System to Reveal Unexpected Interests. In *Proceedings of the International Conference on Multimedia Modelling*, pages 433–438. IEEE, 2005.

Kelly Kaufhold, Sebastian Valenzuela, et Homero Gil De Zúniga. Citizen Journalism and Democracy : How User-generated News Use Relates to Political Knowledge and Participation. *Journalism & Mass Communication Quarterly*, 87(3-4) :515–529, 2010.

Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, et Sanja Fidler. Skip-thought Vectors. In *Proceedings of the Internaltion Conference on Advances in Neural Information Processing Systems*, pages 3294–3302, 2015.

Jon Kleinberg. The Small-world Phenomenon : An Algorithmic Perspective. In *Proceedings of the Conference on Theory of Computing*, pages 163–170. ACM, 2000.

Hans-Peter Kriegel, Peer Kröger, Jörg Sander, et Arthur Zimek. Density-based Clustering. *Wiley Interdisciplinary Reviews : Data Mining and Knowledge Discovery*, 1(3) :231–240, 2011.

G. Krishnalal, S. Babu Rengarajan, et K.G. Srinivasagan. A New Text Mining Approach Based on HMM-SVM for Web News Classification. *International Journal of Computer Applications*, 1(19) :98–104, 2010.

Alex Krizhevsky, Ilya Sutskever, et Geoffrey E Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In *Proceedings of the International Conference on Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.

Robert Krovetz. Homonymy and Polysemy in Information Retrieval. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics*, pages 72–79. Association for Computational Linguistics, 1997.

Haewoon Kwak, Changhyun Lee, Hosung Park, et Sue Moon. What is Twitter, a Social Network or a News Media ? In *Proceedings of the International Conference on World Wide Web*, pages 591–600. ACM, 2010.

Daniel Lamprecht, Markus Strohmaier, et Denis Helic. A Method for Evaluating the Navigability of Recommendation Algorithms. In *Proceedings of the International Workshop on Complex Networks and their Applications*, pages 247–259. Springer, 2016.

Joseph D Lasica. Blogs and Journalism Need Each Other. *Nieman reports*, 57(3) :70–74, 2003.

Hoang An Le, Q.M Bui, Benoît Huet, et et al. LinkedTV at MediaEval 2014 Search and Hyperlinking Task. In *Proceedings of the MediaEval Workshop*, 2014.

Chin-Yew Lin et Eduard Hovy. The Automated Acquisition of Topic Signatures for Text Summarization. In *Proceedings of the International Conference on Computational Linguistics*, pages 495–501. Association for Computational Linguistics, 2000.

Wilson Lowrey. Mapping the Journalism–blogging Relationship. *Journalism*, 7(4) :477–500, 2006.

Minh-Thang Luong, Hieu Pham, et Christopher D Manning. Effective Approaches to Attention-based Neural Machine Translation. In *Proceedings of the International Conference on Empirical Methods in Natural Language Processing*, pages 1412—1421. ACL, 2015.

Juha Makkonen. Investigations on Event Evolution in TDT. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 43–48. Association for Computational Linguistics, 2003.

Christopher D Manning. Computational Linguistics and Deep Learning. *Computational Linguistics*, 2016.

Michael McCandless, Erik Hatcher, et Otis Gospodnetic. *Lucene in Action : Covers Apache Lucene 3.0*. Manning Publications Co., 2010.

Kathleen McKeown, Rebecca J Passonneau, David K Elson, Ani Nenkova, et Julia Hirschberg. Do Summaries Help ? In *Proceedings of the International Conference on Research and Development in Information Retrieval*, pages 210–217. ACM, 2005.

Kathleen R McKeown, Regina Barzilay, David Evans, Vasileios Hatzivassiloglou, Judith L Klavans, Ani Nenkova, Carl Sable, Barry Schiffman, et Sergey Sigelman. Tracking and Summarizing News on a Daily Basis with Columbia's Newsblaster. In *Proceedings of the International Conference on Human Language Technology Research*, pages 280–285. Morgan Kaufmann Publishers Inc., 2002.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, et Jeff Dean. Distributed Representations of Words and Phrases and their Compositionality. In *Proceedings of the Internaltion Conference on Advances in Neural Information Processing Systems*, pages 3111–3119, 2013.

Stanley Milgram. The Small World Problem. *Psychology today*, 2(1) :60–67, 1967.

George A Miller et Walter G Charles. Contextual Correlates of Semantic Similarity. *Language and cognitive processes*, 6(1) :1–28, 1991.

Jeroen Morang, Roeland Ordelman, Franciska de Jong, et Arjan van Hessen. InfoLink : Analysis of Dutch Broadcast News and Cross-media Browsing. In *Proceedings of the International Conference on Multimedia and Expo*, pages 1582–1585. IEEE, 2005.

Masaki Mori, Takao Miura, et Isamu Shioya. Topic Detection and Tracking for News Web Pages. In *Proceedings of the International Conference on Web Intelligence*, WI '06, pages 338–342, Washington, DC, USA, 2006. IEEE Computer Society. ISBN 0-7695-2747-7. doi: 10.1109/WI.2006.171. URL http://dx.doi.org/10.1109/WI.2006.171.

Marius Muja et David G Lowe. Fast Approximate Nearest Neighbors with Automatic Algorithm Configuration. *VISAPP*, 2(331-340) :2, 2009.

Philippe Muller et Xavier Tannier. Annotating and Measuring Temporal Relations in Texts. In *Proceedings of the International Conference on Computational Linguistics*, pages 50–56. ACL, 2004.

Ramesh Nallapati, Ao Feng, Fuchun Peng, et James Allan. Event Threading within News Topics. In *Proceedings of the International Conference on Information and Knowledge Management*, pages 446–453. ACM, 2004.

Hidetsugu Nanba, Noriko Kando, et Manabu Okumura. Classification of Research Papers Using Citation Links and Citation Types : Towards Automatic Review Article Generation. *Advances in Classification Research Online*, 11(1) :117–134, 2011.

Vivi Nastase, Rada Mihalcea, et Dragomir R Radev. A Survey of Graphs in Natural Language Processing. *Natural Language Engineering*, 21(5) :665–698, 2015.

Érik Neuveu et Louis Quéré. Présentation. *Réseaux*, pages 7–21, 1996.

Nic Newman, David A. L. Levy, et Rasmus Kleis Nielsen. Reuters Institute Digital News Report, 2015.

Nic Newman, David A. L. Levy, et Rasmus Kleis Nielsen. Reuters Institute Digital News Report, 2016.

Jakob Nielsen et Thomas K Landauer. A Mathematical Model of the Finding of Usability Problems. In *Proceedings of the INTERACT and CHI Conference on Human Factors in Computing Systems*, pages 206–213. ACM, 1993.

Roeland JF Ordelman, Maria Eskevich, Robin Aly, Benoit Huet, et Gareth Jones. Defining and Evaluating Video Hyperlinking for Navigating Multimedia Archives. In *Proceedings of the International Conference on World Wide Web*, pages 727–732. ACM, 2015.

Paul Over, George Awad, Martial Michel, Jonathan Fiscus, Wessel Kraaij, Alan F. Smeaton, Georges Quénot, et Roeland Ordelman. TRECVID 2015 – An Overview of the Goals, Tasks, Data, Evaluation Mechanisms and Metrics. In *Proceedings of the TRECVID Workshop*, 2015.

Lei Pang et Chong-Wah Ngo. VIREO @ TRECVID 2015 : Video Hyperlinking. In *Proceedings of the TRECVID Workshop*, 2015.

Marcus Jerome Pickering, Lawrence Wong, et Stefan M Rüger. ANSES : Summarisation of News Video. In *Proceedings of the Conference on Image and Video Retrieval*, volume 2728, pages 425–434. Springer, 2003.

Iannis Pledel. Les Blogs, les Promesses d'un Média à Travers ses Représentations Collectives : Illusions ou Réalités à Portée de Clic ? *Les Cahiers du Journalisme*, (16) :252–274, 2006.

Miloš Radovanović, Alexandros Nanopoulos, et Mirjana Ivanović. Hubs in Space : Popular Nearest Neighbors in High-dimensional Data. *Journal of Machine Learning Research*, 11(Sep) :2487–2531, 2010.

Lev Ratinov et Dan Roth. Design Challenges and Misconceptions in Named Entity Recognition. In *Proceedings of the Conference on Computational Natural Language Learning*, pages 147–155. Association for Computational Linguistics, 2009.

Earl Rennison. Galaxy of News : An Approach to Visualizing and Understanding Expansive News Landscapes. In *Proceedings of the Symposium on User Interface Software and Technology*, pages 3–12. ACM, 1994.

S Rasoul Safavian et David Landgrebe. A Survey of Decision Tree Classifier Methodology. *IEEE Transactions on Systems, Man, and Cybernetics*, 21(3) :660–674, 1991.

Gerard Salton et Christopher Buckley. Term-weighting Approaches in Automatic Text Retrieval. *Information processing & management*, 24(5) :513–523, 1988.

Stan Salvador et Philip Chan. Determining the Number of Clusters/Segments in Hierarchical Clustering/Segmentation Algorithms. In *Proceedings of the International Conference on Tools with Artificial Intelligence*, pages 576–584. IEEE, 2004.

Tim Schlippe, Lukasz Gren, Ngoc Thang Vu, et Tanja Schultz. Unsupervised Language Model Adaptation for Automatic Speech Recognition of Broadcast News Using Web 2.0. In *Proceedings of the International Conference Interspeech*, pages 2698–2702, 2013.

David Sculley. Web-scale k-means Clustering. In *Proceedings of the International Conference on World Wide Web*, pages 1177–1178. ACM, 2010.

Klaus Seyerlehner, Peter Knees, Dominik Schnitzer, et Gerhard Widmer. Browsing Music Recommendation Networks. In *Proceedings of the International Conference on Music Information Retrieval*, pages 129–134, 2009.

Dafna Shahaf et Carlos Guestrin. Connecting the Dots Between News Articles. In *Proceedings of the International Conference on Knowledge Discovery and Data mining*, pages 623–632. ACM, 2010.

Tony Silvia. *Global News : Perspectives on the Information Age*. Iowa State Press, 2001.

Anca-Roxana Simon. *Semantic Structuring of Video Collections from Speech : Segmentation and Hyperlinking*. PhD thesis, 2015.

Anca-Roxana Simon, Ronan Sicre, Rémi Bois, Guillaume Gravier, et Pascale Sébillot. IRISA at TRECVid2015 : Leveraging Multimodal LDA for Video Hyperlinking. In *Proceedings of the TRECVID Workshop*, 2015.

Josef Sivic, Bryan C Russell, Alexei A Efros, Andrew Zisserman, et William T Freeman. Discovering Objects and their Location in Images. In *Proceedings of the International Conference on Computer Vision*, volume 1, pages 370–377. IEEE, 2005.

Alan F Smeaton, Noel Murphy, Noel E O'Connor, Sean Marlow, Hyowon Lee, Kieran McDonald, Paul Browne, et Jiamin Ye. The Físchlár Digital Video System : a Digital Library of Broadcast TV Programmes. In *Proceedings of the International Conference on Digital libraries*, pages 312–313. ACM, 2001.

Alan F Smeaton, Cathal Gurrin, Hyowon Lee, Kieran McDonald, Noel Murphy, Noel E O'Connor, Derry O'Sullivan, Barry Smyth, et David Wilson. The Físchlár-news-stories System : Personalised Access to an Archive of TV News. In *Proceedings of the International Conference on Coupling Approaches, Coupling Media and Coupling Languages for Information Retrieval*, pages 3–17, 2004.

Pascal Soucy et Guy W Mineau. A Simple KNN Algorithm for Text Categorization. In *Proceedings of the International Conference on Data Mining*, pages 647–648. IEEE, 2001.

Mirco Speretta et Susan Gauch. Personalized Search Based on User Search Histories. In *Proceedings of the International Conference on Web Intelligence*, pages 622–628. IEEE, 2005.

Mark Steyvers et Tom Griffiths. Probabilistic Topic Models. *Handbook of Latent Semantic Analysis*, 427(7) :424–440, 2007.

Alexander Strehl, Joydeep Ghosh, et Raymond Mooney. Impact of Similarity Measures on Web-page Clustering. In *Proceedings of the International Workshop on Artificial Intelligence for Web Search*, volume 58, page 64, 2000.

Amanda Sturgill, Ryan Pierce, et Yiliu Wang. Online News Websites : How Much Content do Young Adults Want. *Journal of Magazine & New Media Research*, 11(2) :1–18, 2010.

Russell Swan et James Allan. Automatic generation of overview timelines. In *Proceedings of the International Conference on Research and Development in Information Retrieval*, pages 49–56, 2000.

Xavier Tannier et Frédéric Vernier. Creation, Visualization and Edition of Timelines for Journalistic Use. pages 16–19, 2016.

Xavier Tannier, Véronique Moriceau, Béatrice Arnulphy, et Ruixin He. Evolution of Event Designation in Media : Preliminary Study. In *Proceedings of the International Conference on Language Resources and Evaluation*, pages 528–531, 2012.

Mike Thelwall. What is this Link Doing Here ? Beginning a Fine-grained Process of Identifying Reasons for Academic Hyperlink Creation. *Information Research*, 8(3), 2003.

opinion & social TNS. Les Habitudes Médiatiques dans l'Union Européenne, 2015.

Naushad UzZaman, Hector Llorens, James Allen, Leon Derczynski, Marc Verhagen, et James Pustejovsky. TempEval-3 : Evaluating Events, Time Expressions, and Temporal Relations. *arXiv preprint arXiv :1206.5333*, 2012.

Willem Robert Van Hage, Véronique Malaisé, Roxane Segers, Laura Hollink, et Guus Schreiber. Design and use of the Simple Event Model (SEM). *Web Semantics : Science, Services and Agents on the World Wide Web*, 9(2) :128–136, 2011.

Saúl Vargas et Pablo Castells. Rank and Relevance in Novelty and Diversity Metrics for Recommender Systems. In *Proceedings of the Conference on Recommender Systems*, pages 109–116. ACM, 2011.

Marc Verhagen, Robert Gaizauskas, Frank Schilder, Mark Hepple, Graham Katz, et James Pustejovsky. Semeval-2007 task 15 : TempEval Temporal Relation Identification. In *Proceedings of the International Workshop on Semantic Evaluations*, pages 75–80. Association for Computational Linguistics, 2007.

Marc Verhagen, Roser Sauri, Tommaso Caselli, et James Pustejovsky. SemEval-2010 task 13 : TempEval-2. In *Proceedings of the International Workshop on Semantic Evaluations*, pages 57–62. Association for Computational Linguistics, 2010.

Vedran Vukotić, Christian Raymond, et Guillaume Gravier. Multimodal and Crossmodal Representation Learning from Textual and Visual Features with Bidirectional Deep Neural Networks for Video Hyperlinking. In *Proceedings of the International Workshop on Vision and Language Integration Meets Multimedia Fusion*, pages 37–44. ACM, 2016.

Ivan Vulic, Wim De Smet, et Marie-Francine Moens. Cross-language Information Retrieval Models Based on Latent Topic Models Trained with Document-aligned Comparable Corpora. *Information Retrieval*, 16(3) :331–368, 2013. doi: 10.1007/s10791-012-9200-5. URL http://dx.doi.org/10.1007/s10791-012-9200-5.

Ivan Vulic, Wim De Smet, Jie Tang, et Marie-Francine Moens. Probabilistic Topic Modeling in Multilingual Settings : An Overview of its Methodology and Applications. *Information Processing Management*, 51(1) :111–147, 2015. doi: 10.1016/j.ipm.2014.08.003. URL http://dx.doi.org/10.1016/j.ipm.2014.08.003.

Howard D Wactlar. New Directions in Video Information Extraction and Summarization. In *Proceedings of the DELOS Workshop*, pages 24–25, 1999.

Howard D Wactlar, Takeo Kanade, Michael A Smith, et Scott M Stevens. Intelligent Access to Digital Video : Informedia Project. *Computer*, 29(5) :46–52, 1996.

Melissa Wall. 'Blogs of war' Weblogs as News. *Journalism*, 6(2) :153–172, 2005.

James Z Wang, Nozha Boujemaa, Alberto Del Bimbo, Donald Geman, Alexander G Hauptmann, et Jelena Tesić. Diversity in Multimedia Information Retrieval Research. In *Proceedings of the International Workshop on Multimedia Information Retrieval*, pages 5–12. ACM, 2006.

Matthew S Weber et Peter Monge. The Flow of Digital News in a Network of Sources, Authorities, and Hubs. *Journal of Communication*, 61(6) :1062–1081, 2011.

Tim Weninger. An Exploration of Submissions and Discussions in Social News : Mining Collective Intelligence of Reddit. *Social Network Analysis and Mining*, 4(173) :1–19, 2014.

Ross Wilkinson et Allan Smeaton. Automatic Link Generation. *ACM Computing Surveys*, 31(4), 1999.

Lars Willnat et David Hugh Weaver. The American Journalist in the Digital Age : Key Findings, 2014.

Bartosz W Wojdynski et Sriram Kalyanaraman. The Three Dimensions of Website Navigability : Explication and Effects. *Journal of the Association for Information Science and Technology*, 67(2) :454–464, 2016.

Rui Yan, Liang Kong, Congrui Huang, Xiaojun Wan, Xiaoming Li, et Yan Zhang. Timeline Generation Through Evolutionary Trans-temporal Summarization. In *Proceedings of the International Conference on Empirical Methods in Natural Language Processing*, pages 433–443. Association for Computational Linguistics, 2011.

Christopher C Yang, Xiaodong Shi, et Chih-Ping Wei. Discovering Event Evolution Graphs from News Corpora. *IEEE Transactions on Systems, Man, and Cybernetics*, 39 (4) :850–863, 2009.

M-S Yang. A survey of fuzzy clustering. *Mathematical and Computer modelling*, 18(11) : 1–16, 1993.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, et Eduard Hovy. Hierarchical Attention Networks for Document Classification. In *Proceedings of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, pages 1480–1489, 2016.

Table des figures

Liste des tableaux

Liste des publications

— Rémi Bois, Guillaume Gravier, Pascale Sébillot, et Emmanuel Morin. Vers une Typologie de Liens entre Contenus Journalistiques. In *Proceedings of the Conference on Traitement Automatique des Langues Naturelles*, pages 515–521, 2015

— Anca-Roxana Şimon, Rémi Bois, Guillaume Gravier, Pascale Sébillot, Emmanuel Morin, et Sien Moens. Hierarchical Topic Models for Language-based Video Hyperlinking. In *Proceedings of the International Workshop on Speech, Language and Audio in Multimedia*, 2015

— Anca-Roxana Simon, Ronan Sicre, Rémi Bois, Guillaume Gravier, et Pascale Sébillot. IRISA at TRECVid2015: Leveraging Multimodal LDA for Video Hyperlinking. In *Proceedings of the TRECVID Workshop*, 2015

— Guillaume Gravier, Martin Ragot, Laurent Amsaleg, Rémi Bois, Grégoire Jadi, Eric Jamet, Laura Monceaux, et Pascale Sébillot. Shaping-Up Multimedia Analytics: Needs and Expectations of Media Professionals. In *Proceedings of the International Conference on Multimedia Modeling , Perspectives on Multimedia Analytics*, 2016

— Rémi Bois, Vedran Vukotić, Ronan Sicre, Christian Raymond, Guillaume Gravier, et Pascale Sébillot. IRISA at TRECVid2016: Crossmodality, Multimodality and Monomodality for Video Hyperlinking. In *Proceedings of the TRECVID Workshop*, 2016

— Rémi Bois, Vedran Vukotić, Anca-Roxana Simon, Ronan Sicre, Christian Raymond, Pascale Sébillot, et Guillaume Gravier. Exploiting Multimodality in Video Hyperlinking to Improve Target Diversity. In *Proceedings of the International Conference on Multimedia Modeling*, pages 185–197. Springer, Cham, 2017c

— Rémi Bois, Guillaume Gravier, Éric Jamet, Emmanuel Morin, Maxime Robert, et Pascale Sébillot. Linking Multimedia Content for Efficient News Browsing. In *Proceedings of the International Conference on Multimedia Retrieval*, pages 301–307, 2017a

— Rémi Bois, Guillaume Gravier, Eric Jamet, Emmanuel Morin, Pascale Sébillot, et Maxime Robert. Language-based Construction of Explorable News Graphs for Journalists. In *Proceedings of the International Workshop: Natural Language Processing meets Journalism*, pages 31–36, 2017b

Résumé

Cette thèse en informatique s'intéresse à la structuration et à l'exploration de collections journalistiques. Elle fait appel à plusieurs domaines de recherches : sciences sociales, à travers l'étude de la production journalistique ; ergonomie ; traitement des langues et la recherche d'information ; multimédia et notamment la recherche d'information multimédia. Une branche de la recherche d'information multimédia, appelée hyperliage, constitue la base sur laquelle cette thèse est construite. L'hyperliage consiste à construire automatiquement des liens entre documents multimédias. Nous étendons ce concept en l'appliquant à l'entièreté d'une collection afin d'obtenir un hypergraphe, et nous intéressons notamment à ses caractéristiques topologiques.

Nous proposons dans cette thèse des améliorations de l'état de l'art selon trois axes principaux : une structuration de collections d'actualités à l'aide de graphes mutlisources et multimodaux fondée sur la création de liens inter-documents, son association à une diversité importante des liens permettant de représenter la grande variété des intérêts que peuvent avoir différents utilisateurs, et enfin l'ajout d'un typage des liens créés permettant d'expliciter la relation existant entre deux documents. Ces différents apports sont renforcés par des études utilisateurs démontrant leurs intérêts respectifs.

Mots clés : Recherche d'information multimédia, traitement automatique des langues, diversité, hypergraphe

Abstract

This thesis studies the structuring and exploration of news collections. While its main focus is on natural language processing and multimedia retrieval, it also deals with social studies through the study of the production of news and ergonomy through the conduct of user tests. The task of hyperlinking, which was recently put forward by the multimedia retrieval community, is at the center of this thesis. Hyperlinking consists in automatically finding relevant links between multimedia segments. We apply this concept to whole news collections, resulting in the creation of an hypergraph, and study the topological properties of the resulting structure.

In this thesis, we provide improvements on the state of the art along three main axes : a structuring of news collections by mean of mutlisources and multimodal graphs based on the creation of inter-document links, its association with a large diversity of links allowing to represent the variety of interests that different users may have, and a typing of the created links in order to make the nature of the relation between two documents explicit. Extensive user studies confirm the interest of the methods developped in this thesis.

Keywords : Multimedia retrieval, natural language processing, diversity, hypergraph