Filtrage et agrégation d'informations vitales relatives à des entités

Rafik Abbes

THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par : *l'Université Toulouse 3 Paul Sabatier (UT3 Paul Sabatier)*

Présentée et soutenue le *11/12/2015* **par :**

Rafik ABBES

Filtrage et agrégation d'informations vitales relatives à des entités

---

JURY

Nathalie Aussenac-Gilles	DR au CNRS, Toulouse	Président
Patrice Bellot	Professeur, Université d'Aix-Marseille	Rapporteur
Sylvie Calabretto	Professeur, INSA de Lyon	Rapporteur
Faiez Gargouri	Professeur, Université de Sfax	Examinateur
Jian-Yun Nie	Professeur, Université de Montréal	Examinateur
Mohand Boughanem	Professeur, Université Toulouse 3	Directeur
Nathalie Hernandez	Maître de conférence, Université Toulouse 2	Co-encadrante
Karen Pinel-Sauvagnat	Maître de conférence, Université Toulouse 3	Co-encadrante

---

École doctorale et spécialité :
MITT : Image, Information, Hypermedia

Unité de Recherche :
Institut de Recherche en Informatique de Toulouse (UMR 5505)

Directeur(s) de Thèse :
Mohand Boughanem, *Nathalie Hernandez* et *Karen Pinel-Sauvagnat*

Rapporteurs :
Patrice Bellot et *Sylvie Calabretto*

v. 2016-01-15 Commentaires, corrections et autres remarques sont les bienvenus à :
rafik.abbes@irit.fr, rafik.abbes@gmail.com.

Remerciements

C'est à Toulouse, la ville rose, que je passe une étape importante de ma vie. Juin 2012, obtention du diplôme de Master ouvrant la porte à cette thèse. Personne ne peut comprendre cette expérience sans vivre et goûter ces instants. L'enseignement du premier cours, la notification du premier papier accepté, la saveur d'un prix du meilleur papier ...

C'est avec beaucoup d'émotion que j'écris ces lignes pour remercier toutes les personnes qui ont fait de cette thèse une belle expérience réussie.

En premier lieu, je souhaite exprimer mes sincères remerciements aux personnes qui ont participé à ce travail. Mon directeur de recherche *Mohand Boughanem* pour sa confiance et la pertinence des remarques et des conseils qu'il m'a prodigués. Mes co-encadrantes *Karen Pinel-Sauvagnat* et *Nathalie Harnandez* pour leur soutien tout au long de ma thèse et pour leurs relectures attentives et suggestions judicieuses.

Je souhaite également remercier mes rapporteurs Monsieur *Patrice Bellot*, Professeur à l'Université d'Aix-Marseille, et Madame *Sylvie Calabretto*, Professeur à l'Institut National des Sciences Appliquées de Lyon, pour avoir pris le temps de lire et évaluer ce travail et pour la précision et le professionnalisme de leurs retours.

Je tiens également à remercier les examinateurs Monsieur *Faiez Gargouri*, Professeur à l'Université de Sfax et Monsieur *Jian-Yun Nie*, Professeur à l'Université de Montréal d'avoir accepté de participer à mon jury de thèse et échangé leurs points de vue.

Je tiens à remercier infiniment le ministère d'enseignement supérieur et de la recherche scientifique en Tunisie de m'avoir accordé la bourse d'études qui m'a permis de réaliser cette thèse. Je souhaite exprimer également toute ma reconnaissance à Monsieur *Mohamed Tmar* et Monsieur *Mohamed Ben Aouicha* avec qui j'ai fait mes premiers pas en recherche.

Mes remerciements vont de même aux membres des équipes SIG, IRIS

et MELODI de l'IRIT. Je tiens à remercier en particulier l'ensemble des occupants du bureau 401 que j'ai pu côtoyer : Firas, Amjed, Faten, Arlind, Baptise, Thomas, Hung et bien entendu Bileeel avec qui j'ai partagé tous les moments de la thèse, aussi bien dans le laboratoire qu'à l'extérieur, mais aussi le jour de nos soutenances :)

Merci également à tous mes amis de l'IRIT qui m'ont apporté leur soutien moral pendant ces années d'études : Meriam, Ameni, Ismail, Mohamed, Imen, Jean-philippe, Hamid, Eya, Liana, ... Sans oublier de remercier tous mes amis en France et en Tunisie, au nom de l'amitié qui nous a réuni, de nos souvenirs inoubliables et en témoignage de ma profonde affection.

Je voudrais remercier mon très cher frère Tarek et son épouse Rania, vous qui m'avez toujours apporté affection, encouragement et soutien moral. Je vous souhaite plein de bonheur et de réussite.

Enfin, et surtout, je remercie *mes très chers parents Ridha & Sihem*, à qui je dois tout ce que je suis et serai. Pour les sacrifices, les encouragements, pour les vœux tant formulés dans vos prières. Je vous dédie ce modeste travail comme infime hommage à tous les sacrifices consentis et en témoignage de mon grand amour, de ma grande estime et mon immense reconnaissance. Soyez fiers, votre fils est désormais docteur. *Que Dieu vous protège, vous prête bonne santé et longue vie.*

Résumé

Aujourd'hui, les bases de connaissances telles que Wikipedia et DBpedia représentent les sources principales pour accéder aux informations disponibles sur une grande variété d'entités (une entité est une *chose* qui peut être distinctement identifiée par exemple une personne, une organisation, un produit, un événement, etc.). Cependant, la mise à jour de ces sources avec des informations nouvelles en rapport avec une entité donnée se fait manuellement par des contributeurs et avec un temps de latence important en particulier si cette entité n'est pas populaire.

Concevoir un système qui analyse les documents dès leur publication sur le Web pour filtrer les informations importantes relatives à des entités pourra sans doute accélérer la mise à jour de ces bases de connaissances. Dans cette thèse, nous nous intéressons au filtrage d'informations pertinentes et nouvelles, appelées vitales, relatives à des entités. Ces travaux rentrent dans le cadre de la recherche d'information mais visent aussi à enrichir les techniques d'ingénierie de connaissances en aidant à la sélection des informations à traiter.

Nous souhaitons répondre principalement aux deux problématiques suivantes : (1) Comment détecter si un document est vital (c.à.d qu'il apporte une information pertinente et nouvelle) par rapport à une entité donnée ? et (2) Comment extraire les informations vitales à partir de ces documents qui serviront comme référence pour mettre à jour des bases de connaissances ? Concernant la première problématique, nous avons proposé deux méthodes. La première proposition est totalement supervisée. Elle se base sur un modèle de langue de vitalité. La deuxième proposition mesure la fraîcheur des expressions temporelles contenues dans un document afin de décider de sa vitalité. En ce qui concerne la deuxième problématique relative à l'extraction d'informations vitales à partir des documents vitaux, nous avons proposé une méthode qui sélectionne les phrases comportant potentiellement ces informations vitales, en nous basant sur la présence de mots déclencheurs récupérés automatiquement à partir de la connaissance déjà représentée dans la base de connaissances (comme la description d'entités similaires). L'évaluation des approches proposées a été effectuée dans le cadre de la campagne d'évaluation internationale TREC sur une collection de 1.2 milliard de documents avec différents types d'entités

(personnes, organisations, établissements et événements). Pour les approches de filtrage de documents vitaux, nous avons mené nos expérimentations dans le cadre de la tâche "Knwoledge Base Acceleration (KBA)" pour les années 2013 et 2014. L'exploitation des expressions temporelles dans le document a permis d'obtenir de bons résultats dépassant le meilleur système proposé dans la tâche KBA 2013. Pour évaluer les contributions concernant l'extraction des informations vitales relatives à des entités, nous nous sommes basés sur le cadre expérimental de la tâche "Temporal Summarization (TS)". Nous avons montré que notre approche permet de minimiser le temps de latence des mises à jour de bases de connaissances.

Publications

Articles de revues internationales :

1. **Rafik Abbes**, Karen Pinel-Sauvagnat, Nathalie Hernandez, Mohand Boughanem. When Temporal Expressions Help to Detect Vital Documents Related to An Entity. In *ACM SIGAPP Applied Computing Review*, Volume 15 Issue 3, Pages 49-58, Septembre 2015.

Articles de conférences et workshops internationaux :

1. **Rafik Abbes**, Karen Pinel-Sauvagnat, Nathalie Hernandez, Mohand Boughanem. Leveraging temporal expressions to filter vital documents related to an entity. In *Proceedings of the 30th Annual ACM Symposium on Applied Computing (SAC 2015)*, Pages 1093-1098, 2015 (best paper award).
2. **Rafik Abbes**, Nathalie Hernandez, Karen Pinel-Sauvagnat, Mohand Boughanem. Accelerating the update of knowledge base instances by detecting vital information from a document stream (short paper). In *IEEE/WIC/ACM International Conference on Web Intelligence (to appear)*, 2015.
3. **Rafik Abbes**, Karen Pinel-Sauvagnat, Nathalie Hernandez, Mohand Boughanem. IRIT at TREC KBA 2014. In : *Text REtrieval Conference (TREC 2014)*, Gaithersburg, USA, 18/11/2014-21/11/2014, National Institute of Standards and Technology (NIST), 2014.
4. **Rafik Abbes**, Karen Pinel-Sauvagnat, Nathalie Hernandez, Mohand Boughanem. IRIT at TREC Temporal Summarization 2014. In : *Text REtrieval Conference (TREC 2014)*, Gaithersburg, USA, 18/11/2014-21/11/2014, National Institute of Standards and Technology (NIST), 2014.
5. **Rafik Abbes**, Karen Pinel-Sauvagnat, Nathalie Hernandez, Mohand Boughanem. IRIT at TREC Knowledge Base Acceleration 2013 : Cumulative Citation Recommendation Task. In : *Text REtrieval Conference (TREC 2013)*, Gaithersburg, USA, 19/11/2013-22/11/2013, National Institute of Standards and Technology (NIST), 2013.

Articles de conférences et workshops nationaux :

1. **Rafik Abbes**, Nathalie Hernandez, Karen Pinel-Sauvagnat, Mohand Boughanem. Détection d'informations vitales pour la mise à jour de bases de connaissances. In *Journées Francophones d'Ingénierie des Connaissances (IC 2015)*, Rennes, 29/06/2015-03/07/2015, Association Francaise d'Intelligence Artificielle (AFIA), Pages. 147-158, juillet 2015 (papier nommé).
2. **Rafik Abbes**, Karen Pinel-Sauvagnat, Nathalie Hernandez, Mohand Boughanem. Modèles de langue pour la mise à jour d'un profil d'entité. In *Conférence francophone en Recherche d'Information et Applications (CORIA 2014)*, Nancy, 19/03/2014-21/03/2014, LORIA, Pages. 129-143, mars 2014.
3. **Rafik Abbes**, Arlind Kopliku, Karen Pinel-Sauvagnat, Nathalie Hernandez, Mohand Boughanem. Apport du Web et du Web de Données pour la recherche d'attributs (papier court). In *Conférence francophone en Recherche d'Information et Applications (CORIA 2013)*, Neuchâtel, Suisse, 03/04/2013-05/04/2013, Université de Neuchâtel, Pages. 37-46, avril 2013.

Table des matières

Liste des tableaux

Table des figures

Chapitre 1

Introduction

1.1 Contexte de travail

Nous vivons dans une ère où l'information est omniprésente dans tous les secteurs. Elle est véhiculée à travers différents médias classiques tels que les affiches publicitaires, la presse écrite et parlée, ou encore par des médias modernes tels que l'Internet à travers les différentes applications de messageries ou le Web. Ce dernier est aujourd'hui sans conteste, le moyen de diffusion de l'information le plus important qui n'ait jamais existé. Les informations qui y sont diffusées sont de différentes natures (image, texte, vidéo, ...) et abordent des sujets extrêmement variés. Nous nous intéressons dans ces travaux aux informations diffusées sur le Web qui traitent des entités. Une entité est une "chose" du monde réel distinctement identifiable, par exemple une personne, un lieu, une organisation, un événement, etc.

Parallèlement à l'augmentation rapide des informations publiées sur les entités sur le Web, les besoins en information centrés sur les entités ne cessent d'augmenter. Selon des analyses effectuées par Guo *et al.* (2009), environ 71% des requêtes utilisateurs contiennent des entités nommées (noms de personnes, de lieux, de produits, ...). D'autres analyses effectuées par Pound *et al.* (2010) ont révélé que 52% des requêtes ciblent une ou plusieurs entités comme résultats. Cet intérêt croissant pour les entités a encouragé l'apparition de plusieurs bases de connaissances permettant d'accéder facilement et de façon structurée aux informations disponibles sur une grande variété d'entités. Certaines bases de connaissances se sont limitées à éclairer les utilisateurs sur un domaine précis. Par exemple musicbrainz [1] est spécialisée dans la musique, IMDb [2] dans les films et séries, ou encore GeoNames [3] contient des cartes et des informations sur

1. https ://musicbrainz.org
2. http ://www.imdb.com
3. http ://www.geonames.org/

des lieux. D'autres bases de connaissances comme Wikipedia [4], DBpedia [5] et Freebase [6] sont plus générales. Elles couvrent des entités de domaines différents.

Les entités du monde réel représentées dans une base de connaissances peuvent évoluer : un chercheur peut avoir de nouvelles publications, une entreprise peut augmenter son chiffre d'affaire, etc. L'évolution d'une entité peut avoir un impact sur sa description dans une base de connaissances qui peut devenir soit incomplète, car par exemple les nouvelles publications ne sont pas reportées dans la description d'un chercheur, ou bien obsolète, comme par exemple si le chiffre d'affaire de l'entreprise n'est plus correct. Dans ces deux cas, la mise à jour de la base de connaissances est indispensable.

Aujourd'hui, cette mise à jour se fait dans la plupart des cas de façon manuelle par des éditeurs avec un temps de latence important en particulier pour les entités non populaires. Frank *et al.* (2012) ont indiqué que pour un ensemble d'entités non populaires, le temps de latence médian entre la date à laquelle une page Wikipedia est enrichie par une citation d'un document du Web, et la date à laquelle ce document est publié sur le Web est de 365 jours.

L'extraction de connaissances à partir des documents textuels est une approche couramment utilisée pour la constitution de base de connaissances (Petasis *et al.*, 2011). Ces approches reposent souvent sur l'hypothèse selon laquelle le corpus à partir duquel la connaissance est extraite est identifié, que ce soit à partir des pages Wikipedia dans le cas de DBpedia, ou constitué manuellement (Augenstein *et al.*, 2012; Exner et Nugues, 2012). Dans le contexte de la mise à jour de bases de connaissances, la tâche d'identification des textes d'où extraire la connaissance n'est pas triviale. D'une part, certaines approches d'extraction de connaissance à partir du texte analysent les documents dans leur intégralité, or la connaissance sur une entité donnée est souvent décrite uniquement dans quelques phrases du document. D'autre part, lorsqu'on considère en particulier certaines entités comme des entités de type événement (catastrophe naturelle, ...) dont la connaissance établie peut évoluer fréquemment au cours de la période englobant la date de l'événement, les textes sur lesquels ces approches sont appliquées doivent reporter des informations nouvelles et à jour.

Accélérer la mise à jour des bases de connaissances est une problématique actuelle dont le premier enjeu est d'identifier un besoin d'évolution. L'analyse de documents d'où extraire la connaissance à mettre à jour

4. https ://fr.wikipedia.org
5. http ://dbpedia.org/
6. https ://www.freebase.com/

est une solution pour identifier ce besoin (Zablith *et al.*, 2015). La phase d'identification de ces documents est souvent laissée aux concepteurs de la base dans les travaux d'ingénierie de connaissances. Cependant lorsque la base de connaissances comporte des entités largement mentionnées sur le Web, il est dommage de ne pas tirer profit de ces informations.

Concevoir un système qui analyse les documents dès leur publication sur le Web, puis filtre et identifie automatiquement de nouvelles informations pertinentes sur les entités pourra sans doute accélérer la mise à jour de ces bases de connaissances. Un tel système peut être utile non seulement pour la mise à jour des bases de connaissances, mais également pour une variété d'applications ayant pour objectif la détection d'informations nouvelles relatives à un sujet, un événement, etc. A titre d'exemple, lors d'événements importants tels qu'un tournoi mondial ou durant des événements de crise tels qu'une catastrophe naturelle, les utilisateurs peuvent souhaiter être notifiés, à chaque instant, de nouvelles informations relatives à ces événements. Ils peuvent aussi vouloir suivre l'actualité de leurs célébrités préférées.

Les travaux de cette thèse visent à filtrer à partir des documents Web les informations pertinentes et nouvelles relatives à des entités d'intérêt. Ces informations, appelées aussi *informations vitales* doivent être filtrées rapidement, c'est-à-dire, dès qu'elles apparaissent dans le Web. Au delà de la détection des informations, notre objectif est de construire pour chaque entité une synthèse, sorte de résumé temporel, qui agrège les informations vitales et non redondantes sur une entité qui sera émise aux utilisateurs intéressés (les éditeurs chargés de la mise à jour des bases de connaissances, ou bien les fans/abonnés de cette entité, ...). Ce résumé est dit temporel car il peut s'enrichir au cours du temps lorsque l'entité correspondante évolue.

1.2 Problématiques du filtrage et de l'agrégation d'informations relatives à des entités

Un système de filtrage et d'agrégation d'informations relatives à des entités doit prendre en compte les exigences des utilisateurs qui cherchent des résumés d'actualité instantanés et courts. La mise en œuvre d'un tel système est une tâche complexe qui soulève deux problématiques essentielles que nous pouvons traiter en deux étapes :

- *Étape 1 :* Détection des **documents vitaux** à partir d'un flux de documents.
- *Étape 2 :* Construction d'une synthèse sur l'entité en sélectionnant et agrégeant les phrases reportant des informations vitales et non redondantes à partir des documents vitaux. Nous appelons cette

synthèse ***résumé temporel***.

La **notion de vitalité** a été définie par Frank *et al.* (2013, 2014) dans la tâche *Cumulative Citation Recommendation*, proposée dans la piste *Knowledge Base Acceleration* de la campagne *Text Retrieval Conference (TREC)*. Un document est considéré vital s'il reporte une information à la fois pertinente et nouvelle (d'actualité) sur l'entité au moment où il arrive dans le flux. Recommandé à un éditeur d'une base de connaissances, ce document motive la mise à jour de la description actuelle de l'entité.

Concernant la ***première étape*** (détection des documents vitaux), la problématique principale porte sur la détection de la vitalité d'un document. Cette problématique n'est pas triviale pour plusieurs raisons.

- D'abord, un document est considéré comme vital même si cette information est reportée dans une seule de ses phrases, et que le reste de son contenu traite d'autres sujets.
- De plus, un système de filtrage doit être capable de gérer un grand nombre de documents qui apparaissent dans le Web à chaque instant. Ces documents doivent être indexés et analysés dès leur arrivée afin de décider, à la volée, de leur vitalité.

Pour notre part, pour cette première problématique, nous nous sommes particulièrement intéressés aux questions de recherche suivantes :

1. Quelles sont les informations à fournir en entrée au système de filtrage de documents vitaux, pour décrire une entité d'intérêt ? Est ce que le nom d'une entité est suffisant pour décrire l'entité, ou doit-on en fournir une description plus riche (par exemple la page Wikipedia de l'entité) ?
2. La seconde question de recherche porte sur la modélisation de la vitalité. Plusieurs modèles de recherche d'information, tels que les modèles de langue (Ponte et Croft, 1998) et le modèle de pertinence (Lavrenko et Croft, 2001), ont été proposés dans la littérature pour évaluer la pertinence d'un document dans une collection statique de documents.

   Les questions que nous nous posons ici sont les suivantes :

   (a) Est-ce que ces modèles sont également capables d'évaluer la vitalité d'un document dans un flux dynamique ?

   (b) Peut-on et comment estimer un modèle de vitalité par analogie au modèle de pertinence ?

   (c) Est-il nécessaire de considérer tout le contenu d'un document afin d'estimer un modèle de vitalité, ou bien une seule partie (phrases, paragraphes) est-elle suffisante ?

(d) Est-ce que la vitalité est dépendante de l'entité, ou bien peut-elle être estimée de façon globale ?

3. La troisième question de recherche porte sur l'exploitation de certains facteurs comme la fraîcheur pour la détection de la vitalité. La question que nous nous posons ici est la suivante : Est-ce que l'estimation de la fraîcheur en exploitant les expressions temporelles comme `aujourd'hui`, `hier`, `cette semaine`, etc. reportées au sein d'un document, peut aider à la détection de sa vitalité ?

Concernant la ***deuxième étape*** (construction d'une synthèse autour de l'entité), la problématique de recherche principale concerne le filtrage et l'agrégation des informations importantes à partir des documents vitaux détectés et l'identification de celles qui sont nouvelles pour éviter de submerger l'utilisateur avec la même information. Le but ici est de construire en temps réel, un **résumé temporel** court avec des phrases vitales, non redondantes et couvrant les différents aspects de l'entité considérée.

Pour notre part, nous nous sommes intéressés à ces questions :

1. Quelles sont les spécificités d'une phrase vitale ?
2. Existe-t-il des termes pouvant déclencher la détection de sa vitalité ?
3. Est-ce que la détection de la nouveauté basée sur la divergence textuelle, généralement utilisée dans la littérature (Zhang *et al.*, 2002b; Larkey *et al.*, 2002; Kazawa *et al.*, 2002; Liu *et al.*, 2013a; Xi *et al.*, 2013), peut être améliorée en exploitant la présence de nouvelles entités identifiées dans les phrases ?

1.3 Contribution

Nos travaux visent à améliorer le processus de filtrage et d'agrégation d'informations relatives à des entités d'intérêt. Nous nous intéressons aux deux étapes présentées dans la section précédente.

- Concernant la ***première étape***, nous proposons deux méthodes pour filtrer les documents vitaux.

  1. La première méthode est *supervisée*, elle est basée sur les modèles de langue (Ponte et Croft, 1998; Lavrenko et Croft, 2001). Plus précisément, nous proposons d'estimer un modèle de langue reportant la vitalité, par analogie au modèle de pertinence (Lavrenko et Croft, 2001), à partir des documents vitaux. Comme pour l'estimation de la notion de pertinence, ces modèles ont montré de bonnes performances pour mesurer la notion de vitalité particulièrement lorsqu'ils sont estimés de

façon spécifique à chaque entité en considérant tout le contenu des documents (Abbes *et al.*, 2014b,a).

2. La seconde méthode est *non supervisée*. Nous proposons d'estimer la vitalité en exploitant le facteur de fraîcheur estimé à partir des expressions temporelles reportées dans le document (Abbes *et al.*, 2015c,d). Les expérimentations menées ont montré que la fraîcheur représente un critère important pour détecter la vitalité surtout lorsqu'il est estimé en considérant les phrases ou les paragraphes mentionnant l'entité. En effet, plus les dates reportées dans ces parties sont proches de la date de publication du document, meilleure est la probabilité de vitalité du document.

Afin d'évaluer nos approches de détection de documents vitaux, nous nous sommes basés sur les corpus de documents fournis par la campagne d'évaluation TREC dans la tâche *Knowledge Base Acceleration* en 2013 et 2014.

- Pour la ***deuxième étape***, nous proposons deux solutions pour résoudre les deux problématiques relatives à la *sélection des phrases vitales* et la *détection de la nouveauté (non redondance)* (Abbes *et al.*, 2015b,a).
  - D'abord, nous proposons de sélectionner les phrases vitales en nous basant sur la présence de mots importants (déclencheurs) récupérés automatiquement en exploitant le vocabulaire propre aux entités similaires ayant le même type que l'entité considérée dans une base de connaissances (dans notre cas, *DBpedia*). Les expérimentations menées ont montré que ces mots déclencheurs permettent de couvrir différents aspects de l'entité (stabilité du rappel) et améliorent la sélection des phrases vitales (amélioration de la précision).
  - Ensuite, nous proposons une méthode qui détecte la nouveauté (non redondance) en combinant la divergence textuelle avec l'identification des entités liées dans les phrases. Cette combinaison a montré une légère amélioration dans les performances.

Nous avons combiné ces deux solutions afin de générer un résumé temporel sur l'entité, qui selon les expérimentations, a montré son intérêt pour accélérer la mise à jour des bases de connaissances.

Pour évaluer les méthodes proposées dans cette étape (génération d'un résumé temporel), nous nous sommes basés sur les corpus de documents fournis par la campagne d'évaluation TREC dans la tâche *Temporal Summarization* en 2013 et 2014.

1.4 Organisation du mémoire

Ce mémoire est organisé en deux parties : la première présente le contexte général dans lequel se situe notre travail, à savoir le filtrage d'information centré sur des entités. La seconde partie détaille notre contribution.

L'objectif de la première partie, intitulée ***Synthèse des travaux de l'état de l'art***, est de présenter les concepts de base de la recherche d'information ainsi que ses nouveaux paradigmes centrés sur des entités.

Cette partie est constituée de deux chapitres :

- Le chapitre 2 introduit les notions de base de la recherche et filtrage d'information.
- Le chapitre 3 présente la recherche d'information orientée entités. Nous distinguons ainsi deux axes : le premier axe se focalise sur la recherche d'entités, alors que le second concerne la recherche d'information centrée sur des entités. Les travaux de cette thèse portent sur le deuxième axe et plus spécifiquement sur le filtrage d'informations centrées sur des entités.

La seconde partie de ce mémoire, intitulée ***Filtrage et agrégation d'informations vitales autour d'entités***, présente nos contributions. Cette partie est constituée de deux chapitres :

- Le chapitre 4 décrit nos deux méthodes proposées pour répondre à la première question de recherche, qui concerne la détection des documents vitaux relatifs à des entités. La première méthode se base sur l'utilisation des modèles de langue pour modéliser la vitalité. La deuxième méthode exploite les expressions temporelles reportées dans le document pour estimer sa vitalité.
- Le chapitre 5 expose notre méthode de génération de résumés temporels sur des entités. Cette méthode s'intéresse à deux problématiques : la sélection des phrases vitales et la détection de la nouveauté (non redondance).

L'ensemble des évaluations se basent sur les corpus de documents fournis par la campagne d'évaluation TREC (Text Retrieval Conference) dans les tâches Knowledge Base Acceleration et Temporal Summarization des années 2013 et 2014.

En conclusion, nous dressons le bilan de nos travaux reliés au filtrage d'informations vitales. Nous introduisons ensuite les limites et les perspectives de ces travaux à court et à long terme.

Première partie : Synthèse des travaux de l'état de l'art

Chapitre 2

Recherche et filtrage d'information : concepts de base

2.1 Introduction

L'augmentation exponentielle des données numériques pose de nouveaux défis pour l'accès à l'information qui se fait soit de manière délibérée (recherche d'information *ad-hoc*) à travers un Système de Recherche d'Information (SRI) soit de manière passive à travers un Système de Filtrage d'Information (SFI). Dans ce chapitre, nous présentons les concepts de base de ces deux types d'accès ainsi que leurs principales différences.

2.2 Recherche d'information

La recherche d'information est une branche de l'informatique qui s'intéresse à l'acquisition, l'organisation, le stockage, la recherche et la sélection d'information (Salton et McGill, 1986). L'objectif principal de la recherche d'information est de trouver dans une ou plusieurs collections de documents, ceux qui sont susceptibles de répondre à un besoin d'information utilisateur exprimé à l'aide d'une ou plusieurs requêtes. Pour atteindre cet objectif, plusieurs étapes sont nécessaires. Elles sont modélisées par ce qu'on appelle "Processus en U" de la recherche d'information (Belkin et Croft, 1992).

Le processus de recherche d'information tente d'établir une correspondance entre les besoins en information de l'utilisateur d'une part et les informations disponibles d'autre part. Ce processus, schématisé par la figure 2.2, est composé de trois fonctions principales :

- L'*indexation* des documents et des requêtes de l'utilisateur ;

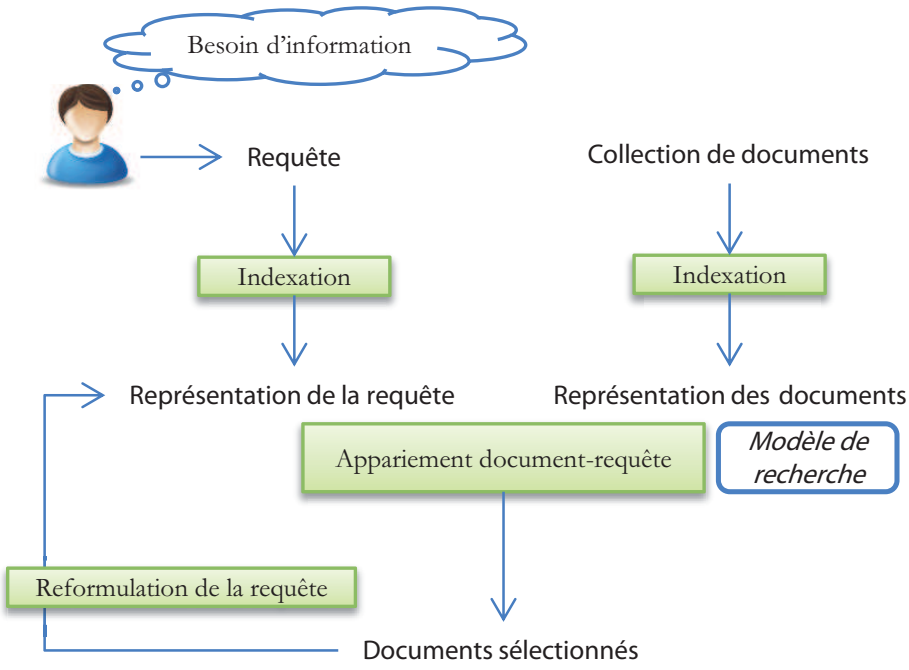

FIGURE 2.1 – Processus en U de la recherche d'information (Belkin et Croft, 1992)

- L'*appariement* document-requête ;
- La *reformulation* de la requête.

Dans la suite, nous détaillons chacune de ces fonctions.

2.2.1 Indexation

Un système de recherche d'information doit renvoyer les informations pertinentes susceptibles de satisfaire le besoin en information de l'utilisateur dans un temps d'exécution acceptable. Afin d'optimiser ce temps, il est nécessaire d'effectuer certaines opérations sur les documents bruts afin de faciliter leur exploitation.

Indexer un document consiste à en extraire un ensemble de mots-clés qui formeront ses descripteurs. Considérer uniquement les mots-clés permet de faciliter l'étape de recherche ultérieurement. L'indexation peut être :

- *Manuelle :* la représentation du document se fait par un spécialiste du domaine (documentaliste) ;

- *Automatique :* le processus d'indexation est entièrement informatisé ;
- *Semi-automatique :* le processus d'indexation se fait automatiquement. Toutefois, le documentaliste peut intervenir pour choisir d'autres termes significatifs et valider la représentation finale du document.

L'indexation automatique (Maron et Kuhns, 1960) regroupe un ensemble de traitements automatisés sur un document comme :

- *l'extraction des mots :* Ce processus consiste à analyser le texte d'un document afin d'extraire ses mots en reconnaissant les espaces de séparation des mots, les ponctuations, etc.
- *l'élimination des mots vides :* Un document contient souvent des mots non significatifs appelés *mots vides* (pronoms personnels, prépositions). L'élimination de ces mots se fait à l'aide d'une liste prédéfinie de mots vides (appelée *stop-list*) ou en supprimant les mots ayant une fréquence dépassant un certain seuil. Éliminer les mots vides permet de réduire la taille de l'index, gagner en espace mémoire et optimiser le temps d'exécution. Cependant, la performance d'un SRI peut baisser notamment lorsque la requête contient une entité nommée avec des mots vides comme `"C à vous"` (émission de télévision diffusée sur `France 5`).
- *la lemmatisation :* Ce traitement consiste à radicaliser les mots restants, c'est à dire réduire les mots à leur forme canonique. Grâce à la lemmatisation, les documents contenant différentes formes d'un même terme auront les mêmes chances d'être restitués ce qui améliore la capacité d'un SRI à retrouver les documents pertinents. Parmi les méthodes utilisées pour la lemmatisation on peut citer l'algorithme de Porter (Porter, 1997) pour les textes en anglais et la troncature (Mayfield et McNamee, 2003) pour les autres langues (Français, Italien, Allemand).
- *la pondération :* Les termes d'un document n'ont pas souvent la même importance. Un terme qui apparaît dans la majorité des documents de la collection aura moins d'importance qu'un terme qui existe dans quelques documents seulement. Plusieurs fonctions de pondération de termes ont été proposées dans la littérature. La plupart de ces fonctions combinent des variantes des facteurs *TF (Term Frequency)* et *IDF (Inverse Document Frequency)* qui mesurent un poids local (dans le document) et global (dans la collection) d'un terme (Manning *et al.*, 2008, chapitre 6).

2.2.2 Appariement Document-Requête

La requête utilisateur devra être comparée aux représentations des documents. Un score de similarité mesurant la pertinence système est calculé pour chaque couple *document-requête* permettant ainsi de restituer

à l'utilisateur une liste ordonnée de documents. L'appariement *document-requête* repose sur un cadre théorique défini par un modèle de recherche d'information. Plusieurs modèles de recherche d'information ont été proposés dans la littérature. Baeza-Yates et Ribeiro-Neto (2011) ont classifié ces modèles en trois catégories :

- *Les modèles* ***ensemblistes*** *:* qui se fondent sur la théorie des ensembles. On distingue le modèle booléen pur (*Boolean Model*) (Salton, 1968), le modèle booléen étendu (*Extended Boolean Model*) (Salton et McGill, 1986) et le modèle basé sur les ensembles flous (*Fuzzy Set Model*) (Ogawa *et al.*, 1991).
- Les *modèles* ***vectoriels*** *:* qui se basent sur l'algèbre, plus précisément le calcul vectoriel. Ils englobent le modèle vectoriel (*Vector Space Model*) (Salton *et al.*, 1975), le modèle vectoriel généralisé (*Generalized Vector Spaces Model*) (Wong *et al.*, 1985), *Latent Semantic Indexing* (*LSI*) (Furnas *et al.*, 1988) et le modèle connexionniste (Boughanem, 1992).
- Les *modèles* ***probabilistes***, qui se basent sur les probabilités. Ils comprennent le modèle probabiliste général (Robertson *et al.*, 1994), le modèle de réseau d'inférence (Turtle et Croft, 1990) et les modèles de langue (Ponte et Croft, 1998; Boughanem *et al.*, 2004).

Nous présentons dans la suite les principaux modèles issus de chacune de ces trois classes. Nous renvoyons le lecteur aux nombreux manuels introductifs à la recherche d'information pour des présentations exhaustives des modèles de RI (Baeza-Yates et Ribeiro-Neto, 2011; Manning *et al.*, 2008; Chowdhury, 2010).

2.2.2.1 Modèle booléen

Le modèle booléen (Salton, 1968) est le modèle le plus ancien dans la recherche d'information. Il est basé sur la théorie des ensembles et l'algèbre de Boole. Dans ce modèle, un document est représenté par un ensemble de termes. Une requête est représentée sous forme d'une expression logique composée de termes reliés par des opérateurs logiques ***ET***, ***OU***, ***NON***. Le modèle booléen repose sur un appariement exact qui consiste à ne restituer que les documents répondant exactement à la requête. Par conséquent, le score de similarité entre un document d et une requête q est inclus dans l'ensemble $\{0, 1\}$.

Le modèle booléen affirme que chaque document est soit pertinent soit non-pertinent, donc répond exactement à la requête qui a été formulée. Il n'y a pas de notion de réponse partielle aux conditions de la requête ce qui implique la restitution d'un nombre très important ou très faible de documents non ordonnés. D'autre part, il est souvent très difficile pour l'utilisateur d'exprimer son besoin en information avec des expressions

booléennes ce qui ne permet pas d'utiliser au mieux les caractéristiques de ce modèle. Enfin, tous les termes dans un document ou dans une requête sont pondérés de la même façon simple (0 ou 1), donc tous les termes ont la même importance pour le document. Ces inconvénients nous amènent à introduire le modèle vectoriel.

2.2.2.2 Modèle vectoriel

Le modèle vectoriel (nommée aussi *VSM* pour *Vector Space Model*) est un modèle statistique proposé par Salton *et al.* (1975). Dans ce modèle les documents et les requêtes sont représentés par des vecteurs d'indexation dans un espace Euclidien de dimension élevée engendré par les termes d'indexation.

Considérons que l'index comporte n termes, un document d_j est représenté par un vecteur $\vec{d_j} = \{\omega_{1,j},\ \omega_{2,j}, ..., \omega_{n,j}\}$, tel que $\omega_{i,j}$ est le poids du terme i dans le document d_j. Dans ce même espace vectoriel, considérons la requête q représentée par un vecteur $\vec{q} = \{\omega_{1,q},\ \omega_{2,q},\ ..., \omega_{n,q}\}$ où $\omega_{i,q}$ est le poids du terme i dans la requête q (généralement 0 ou 1 selon que le terme appartient ou pas à la requête).

La similarité entre le document d_j et la requête q peut être calculée par le cosinus de l'angle $(\vec{d_j}, \vec{q})$ formé par les deux vecteurs :

$$\begin{aligned} score(d_j, q) &= cos(\vec{d_j}, \vec{q}) \\ &= \frac{\vec{d_j}.\vec{q}}{|\vec{d_j}|.|\vec{q}|} \\ &= \frac{\sum_{i=1}^{n} \omega_{i,j}.\omega_{i,q}}{\sqrt{\sum_{i=1}^{n} \omega_{i,j}^2 . \sum_{i=1}^{n} \omega_{i,q}^2}} \end{aligned} \tag{2.1}$$

Le modèle vectoriel permet d'établir un appariement partiel ou approximatif entre les documents et la requête. Il permet d'ordonner les documents selon leur degré de similarité vis-à-vis de la requête. La longueur des documents peut être traitée naturellement lors de l'appariement, car elle est considérée dans le calcul des poids des termes. Le modèle vectoriel est populaire en recherche d'information malgré son inconvénient majeur qui consiste à supposer que les termes de l'index sont indépendants, alors que les termes dans les documents sont souvent liés sémantiquement.

2.2.2.3 Modèle de langue

A la différence du modèle vectoriel qui se base sur une intuition mathématique-géométrique, le modèle de langue (Ponte et Croft, 1998) se base sur la théorie des probabilités. Le modèle de langue permet de modéliser

un texte (un document ou une collection de documents) en capturant la distribution de chacun de ses mots. Soit θ_d un modèle de langue estimé à partir d'un document d. A partir de ce modèle de langue θ_d, il est possible de calculer la probabilité d'observer un terme t ou une séquence de termes $s = t_1 t_2 ... t_n$ dans ce document. Cette probabilité est notée par $P(t|\theta_d)$. Le modèle de langue est dit *unigramme* si la séquence s est composée d'un seul terme, *bigramme* si la séquence s contient deux termes et *n-grammes* si la séquence s contient n termes.

L'estimation de la probabilité d'observer un terme t dans un document d est généralement effectuée par l'estimation du Maximum de vraisemblance (*Maximum likelihood*) comme suit :

$$P(t|\theta_d) = \frac{freq(t,d)}{\sum_{t' \in d} freq(t',d)} \tag{2.2}$$

$freq(t,d)$ est le nombre d'occurrence du terme t dans le document d.

Le modèle de langue est basé sur l'hypothèse que l'utilisateur fournit sa requête en pensant à un ou plusieurs documents qu'il souhaite retrouver. La requête utilisateur est alors inférée à partir des documents. Évaluer la pertinence d'un document par rapport à une requête revient à estimer la probabilité que la requête soit inférée par le document (Ponte et Croft, 1998; Boughanem *et al.*, 2004).

Considérons une requête utilisateur q composée de termes t, la pertinence d'un document d vis-à-vis la requête q est mesurée par la probabilité que la requête puisse être générée par le modèle de langue du document θ_d. Cette probabilité est notée par $P(q|\theta_d)$.

En se basant sur le principe d'indépendance des termes (l'apparition d'un terme n'influe pas la probabilité d'apparition d'un autre terme), la probabilité $P(q|\theta_d)$ peut être écrite en :

$$P(q|d) = \prod_{t \in q} P(t|\theta_d) \tag{2.3}$$

L'estimation de la probabilité $P(t|d)$ par le principe du Maximum de vraisemblance peut conduire à une probabilité nulle dans le cas ou un terme t est absent dans le document d. Pour remédier à ce problème, plusieurs techniques de lissages (*smoothing*) (Zhai et Lafferty, 2004) sont proposées comme le lissage de Laplace, le lissage de Good-Turing (Ney *et al.*, 1995), le lissage de Backoff (Katz, 1987), le lissage par interpolation de *Jelinek-Mercer* (Jelinek et Mercer, 1980) et le lissage de Dirichlet (MacKay et Peto, 1994).

Lavrenko et Croft (2001) ont proposé de modéliser la notion de pertinence de manière explicite. L'idée consiste à considérer que pour chaque requête, il existe un modèle, qu'ils appellent un modèle de pertinence, permettant de générer le sujet abordé par la requête.

Comme la pertinence est inconnue a priori, les auteurs ont proposé d'exploiter les documents les mieux classés pour la requête considérée. Formellement, soit q une requête, la probabilité de générer un terme t à partir du modèle de pertinence θ_R correspondant au sujet de la requête q est exprimée comme suit :

$$P(t|\theta_R) = \sum_{d \in R} P(t|\theta_d) P(q|\theta_d) \tag{2.4}$$

Où R dénote l'ensemble des documents mieux classés pour la requête q.

2.2.3 Reformulation de la requête

Il est souvent difficile de retrouver des informations pertinentes en utilisant la requête initiale de l'utilisateur. Il peut manquer certains termes pertinents ou le vocabulaire employé par l'utilisateur peut ne pas correspondre à celui employé dans les documents. Afin de faire correspondre au mieux la pertinence utilisateur et la pertinence du système, une étape de reformulation de la requête est souvent considérée. Cette étape consiste à modifier la requête au fur et à mesure de la session de recherche par l'ajout de termes significatifs et/ou ajustement de leurs poids.

La reformulation de la requête est effectuée soit manuellement par l'utilisateur qui soumet lui même une nouvelle requête, soit automatiquement en construisant une nouvelle requête basée sur les jugements de pertinence de l'utilisateur (Salton, 1971), ou bien encore en se basant sur un modèle qui exploite les n premiers documents renvoyés (Cao *et al.*, 2008; Xu *et al.*, 2009).

2.2.4 Évaluation

Évaluer adéquatement un système de recherche d'information demeure l'un des principaux défis de ce domaine. L'intérêt est de pouvoir comparer les SRI entre eux selon des critères objectifs qui représentent les attentes de l'utilisateur sur les résultats renvoyés (pertinence, diversité, spécificité, temps de réponse, nouveauté, fraîcheur, etc.)

Une des évaluations la plus courante consiste à créer un environnement unique permettant de comparer les systèmes équitablement (évaluation selon le paradigme de Cranfield (Sanderson, 2010; Saracevic, 1995)). Cet environnement comporte trois parties essentielles :

- Les requêtes (ou *topics*) : décrivant le sujet de la recherche (*quoi ?*).

- Le corpus de documents : l'ensemble de documents présélectionnés présentant le champ de la recherche (*où ?*).
- Les jugements de pertinence (appelé aussi *vérité terrain*, *ground-truth* ou *gold-standard*) : l'ensemble des documents pertinents pour chaque requête. La pertinence d'un document vis-à-vis d'une requête peut être binaire ou graduelle.

2.2.4.1 Campagnes d'évaluation

Plusieurs campagnes d'évaluation sont apparues dans le but d'évaluer des systèmes de recherche d'information sur une tâche bien spécifique. Une tâche est définie par sa description (but, règles, ...), l'ensemble des requêtes d'évaluation et le corpus d'où extraire les réponses. Les jugements de pertinence sont effectués par un ensemble d'individus, appelés juges, qui annotent les documents du corpus. Comme le nombre de documents à annoter peut être grand, une technique de *pooling* peut être appliquée (Jones et van Rijsbergen, 1975). Elle consiste à considérer uniquement les premiers documents retournés par chaque participant afin de constituer le *pool* de documents à juger.

Les campagnes d'évaluation les plus connues sont :

- Campagne TREC (Text REtrieval Conference) [1] : c'est la campagne d'évaluation la plus connue dans le domaine de la recherche d'information. Elle est organisée annuellement depuis 1992 par le NIST (US National Institute of Standards and Technology). TREC a proposé à ce jour un large panel de tâches, telles que *recherche ad-hoc*, *TREC entity*, *TREC Temporal Summarization* et *TREC Knowledge Base Acceleration*
- Campagne INEX (Initiative for the Evaluation of XML Retrieval) [2] : Cette campagne s'est intéressée jusqu'à l'année 2013 à la recherche d'information dans des documents structurés.
- *NTCIR* (NII Testbeds and Community for Information access Research) [3] : Cette campagne s'intéresse principalement à l'accès d'information pour la langue asiatique.
- *CLEF* (Conference and Labs of the Evaluation Forum) : Cette campagne s'est concentrée dans ses premières années sur les langues européennes et la recherche d'information multilingue [4]. Aujourd'hui, elle propose une variété de tâches (labs [5]) de recherche sur des documents (*Question Answering, Social Book Search, News Recom-*

1. http ://trec.nist.gov/
2. http ://inex.mmci.uni-saarland.de/
3. http ://ntcir.nii.ac.jp/
4. http ://www.clef-campaign.org/
5. http ://clef2015.clef-initiative.eu/CLEF2015/lab_overview.php

mandation Evaluation, ...).

Les campagnes d'évaluation évaluent les systèmes participants selon des critères bien définis appelés mesures d'évaluation. Plusieurs mesures ont été proposées dans la littérature dont les principales sont présentées dans la sous-section suivante.

2.2.4.2 Mesures d'évaluation de base

Le but principal des systèmes de recherche d'information est de sélectionner les documents pertinents répondant aux besoins utilisateur et rejeter les documents non pertinents. Pour évaluer la capacité d'un système à atteindre ce but, deux principales mesures sont utilisées : le *rappel* et la *précision*.

Le rappel mesure la capacité du système à retrouver les documents pertinents. Il est calculé par le ratio du nombre de documents pertinents sélectionnés par le système par rapport au nombre de documents pertinents pour la même requête. La précision mesure la capacité du système à rejeter les documents non pertinents. Elle est calculée par le ratio du nombre de documents pertinents sélectionnés par le système par rapport au nombre total de documents sélectionnés pour la même requête.

Formellement, soit R l'ensemble des documents pertinents pour la requête dans la collection et S l'ensemble de documents sélectionnés par le système. Le rappel et la précision sont calculés par les deux équations suivantes :

$$Rappel = \frac{|R \cap S|}{|R|} \tag{2.5}$$

$$Précision = \frac{|R \cap S|}{|S|} \tag{2.6}$$

Pour mesurer la capacité du système à sélectionner le maximum de documents pertinents, et seulement ces documents-ci, une combinaison entre le rappel et la précision peut être utilisée. Cette combinaison, appelée la F_{mesure}, est calculée selon une moyenne harmonique entre le rappel et la précision traduisant le double objectif d'un SRI qui est de minimiser le silence et le bruit (Equation 2.7).

$$F_{mesure} = 2 * \frac{Rappel * Précision}{Rappel + Précision} \tag{2.7}$$

Les mesures de *précision*, du *rappel* et la F_{mesure} sont basées sur des ensembles non ordonnés de documents. Dans un contexte de recherche d'information qui prend en compte le rang des documents, les ensembles appropriés de documents sont naturellement donnés par le top-k premiers documents renvoyés, ce qui revient à calculer la $Précision@k$, $Rappel@k$ et la $F_{mesure@k}$.

Contrairement aux mesures précédentes qui évaluent les documents renvoyés en considérant des jugements de pertinence binaires (pertinents ou non), la mesure de $NDCG@k$ ($NDCG$ au rang k) *Normalized Discounted Cumulative Gain* permet l'évaluation des systèmes de recherche d'information en considérant différents degrés de pertinence. Cette mesure repose sur des jugements de pertinence graduels (Järvelin et Kekäläinen, 2002).

2.3 Filtrage d'information

L'apparition de l'internet et en particulier des technologies du Web 2.0 a complètement révolutionné la manière d'accéder à l'information. De nombreux documents nouveaux sont publiés chaque heure. En effet, si avec un système de recherche d'information classique l'utilisateur accède volontairement à l'information via des requêtes, on assiste aujourd'hui de plus en plus à la prolifération de services qui ramènent ou recommandent des documents susceptibles d'intéresser l'utilisateur. Le processus qui permet de sélectionner l'information désirée à partir d'un flux continu de données (articles d'actualité, posts, emails, ...) s'appelle *filtrage d'information* (Belkin et Croft, 1992).

Un système de filtrage d'information permet d'extraire à partir d'un flux d'informations celles qui sont pertinentes pour des préférences spécifiques représentées par un *profil*. Ce profil décrit les besoins en information permanents (centres d'intérêts de l'utilisateur). Cette description peut correspondre par exemple à un concept général comme *le sport*, ou *le cinéma*, ou bien à une entité plus spécifique comme par exemple *Université Paul Sabatier* ou *Michael Schumacher*, etc.

Le filtrage d'information présente plusieurs problématiques dont les principales sont : la représentation des documents et des profils ainsi que la construction de la fonction permettant de décider si un document donné est pertinent. En effet, en l'absence d'une collection de documents fixe, toutes les techniques de pondération et d'appariement utilisées en recherche d'information ne peuvent pas être utilisées telles quelles (Tmar, 2002).

La figure 2.3 illustre le modèle général d'un système de filtrage d'information qui comporte trois parties :

- Un flux de documents produit des textes tels que des articles de journaux en ligne, des contenus générés par l'utilisateur dans les réseaux sociaux ou dans les forums etc.
- L'utilisateur (ou un groupe d'utilisateurs) précise ses intérêts qui seront représentés par des profils. Le profil représente un besoin en information permanent de l'utilisateur.
- La fonction de décision effectue un appariement entre les documents et

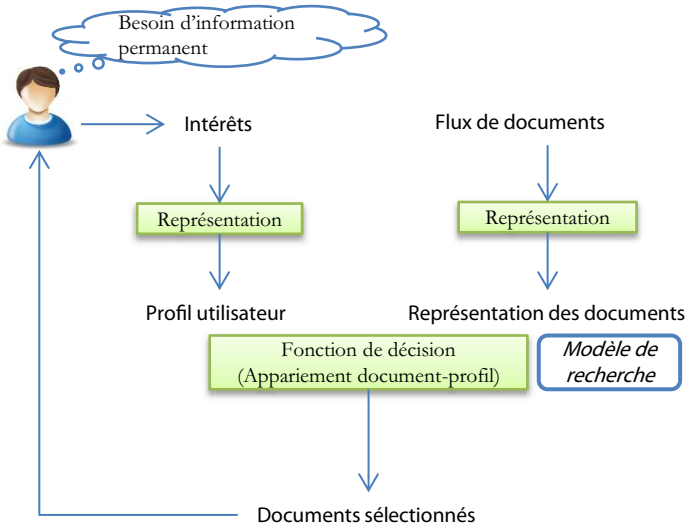

FIGURE 2.2 – Modèle général de filtrage d'information (Belkin et Croft, 1992)

le profil afin de décider l'acceptation ou le rejet d'envoi d'un document vers l'utilisateur.

Un système de filtrage est donc un assistant qui lit et filtre les informations que l'on reçoit et décide à la volée si le document correspond ou pas aux besoins en information des utilisateurs.

Malone *et al.* (1987) a distingué trois formes de filtrage :

- le filtrage social (ou collaboratif) (Malone *et al.*, 1987; Goldberg *et al.*, 1992; Herlocker *et al.*, 1999) : dans cette forme de filtrage les utilisateurs du même système collaborent entre eux. Recevant un document, un utilisateur peut annoter ou recommander ce document à un autre utilisateur. Le système se base sur ces annotations pour générer des règles de sélection permettant d'évaluer un document et le recommander aux autres utilisateurs,
- le filtrage économique (Malone *et al.*, 1987) : il consiste à sélectionner des articles en se basant sur le coût et l'intérêt de produire et de lire les articles. Le but est de sélectionner l'information qui minimise le coût et maximise l'intérêt. Le coût et l'intérêt sont représentés par des paramètres définis par l'utilisateur,
- le filtrage par le contenu (Malone *et al.*, 1987; Lang, 1995; Robertson et Walker, 1999; Tmar, 2002; Tebri, 2004) : ce filtrage se base uniquement sur le contenu du document et du profil de l'utilisateur

pour décider l'acceptation ou le rejet d'un document.

Les systèmes de filtrage peuvent aussi se distinguer par rapport au mode de sélection de l'information. En effet, cette sélection peut se faire de manière synchrone (solution adaptative), c'est à dire déclenchée à chaque arrivée d'un document ou de manière asynchrone (solution *batch*), c'est-à-dire effectuée selon, soit des périodes de temps spécifiques soit sur la base de la quantité de documents reçue.

Nos travaux rentrent dans le cadre de filtrage par le contenu adaptatif. Comme nous le verrons dans la sous-section suivante, ce type de filtrage est souvent vu comme une recherche d'information inversée.

2.3.1 Différences entre la recherche et le filtrage d'information

Le filtrage d'information est une fonction duale de la recherche d'information. Cette dualité est traduite par les faits suivants :

- un système de recherche d'information suppose l'existence d'une collection de documents statique et organisée, alors qu'un système de filtrage d'information traite un flux dynamique de documents diffusés,
- un système de recherche d'information se base sur la requête fournie par l'utilisateur pour répondre à son besoin en information temporaire, au contraire d'un système de filtrage d'information qui vise à répondre à un besoin permanent décrit par un profil,
- un système de recherche d'information cherche les documents intéressants et les ordonne selon une liste classée par ordre de pertinence alors qu'un système de filtrage d'information permet de décider si un document est intéressant ou non,
- un système de filtrage d'information simule un processus peu ou pas interactif, puisque les utilisateurs consultent les documents périodiquement dans le temps. Au contraire, un système de recherche d'information interagit souvent avec l'utilisateur qui consulte les résultats de recherche.

Les systèmes de recherche d'information partagent cependant plusieurs caractéristiques avec les systèmes de filtrage d'information. Les documents et les requêtes sont représentés par des mots clés pondérés. La pondération exploite les fréquences des termes dans le document et la requête/profil. La décision de pertinence mesure une correspondance entre les représentations des documents et des requêtes/profils. Plusieurs travaux ont été entrepris dans ce domaine (Tmar, 2002).

2.3.2 Évaluation des systèmes de filtrage d'information

Les systèmes de filtrage d'information doivent décider pour chaque document du flux s'il est pertinent ou non par rapport à un profil donné. Parfois, le système affecte à un document sélectionné un score de confiance reflétant son degré de pertinence. Selon la présence ou l'absence d'un score de pertinence, nous distinguons deux familles de mesures d'évaluations pour les systèmes de filtrage d'information.

2.3.2.1 Mesures d'évaluation ensemblistes

Avec l'absence d'un score de confiance pour chacun des documents sélectionnés, le résultat de filtrage est un ensemble non ordonné de documents. Les mesures d'évaluation basées sur des listes ordonnées de documents telle que la précision ou le rappel à un rang donné k ne sont pas applicables. Plusieurs mesures alternatives ont été proposées par la campagne d'évaluation TREC pour évaluer les systèmes de filtrage sans tenir compte d'un tri (Hull, 1998; Hull et Robertson, 1999; Robertson et Hull, 2000). Dans de cadre, un document quelconque peut appartenir à l'un de ces quatre ensembles :

- ensemble de documents pertinents sélectionnés (R_+),
- ensemble de documents pertinents rejetés (R_-),
- ensemble de documents non pertinents sélectionnés (N_+),
- ensemble de documents non pertinents rejetés (N_-).

En se basant sur ces ensembles, plusieurs mesures d'évaluation peuvent être appliquées :

Utilité absolue

La mesure de l'utilité absolue (Hull, 1998) d'un système pour un profil donné est la somme linéaire suivante :

$$Utilité\ absolue = w_1 * R_+ + w_2 * R_- + w_3 * N_+ + w_4 * N_- \tag{2.8}$$

$w_i (i \in [1, 4])$ est un paramètre d'utilité qui peut être positif pour donner plus d'importance à la bonne classification d'un document dans une catégorie, ou négatif pour exprimer le coût de la mauvaise classification d'un document dans une catégorie. Par conséquent, plus le score d'utilité est grand, plus la performance du système est grande.

Il est difficile de comparer les systèmes en se basant sur l'utilité absolue. En effet, les scores d'utilité varient largement d'un profil à un autre en fonction du nombre de documents pertinents. Le simple calcul de l'utilité moyenne donne à tous les documents retrouvés le même poids, ce qui signifie que les moyennes d'utilité vont être très influencées par les profils ayant beaucoup de documents pertinents.

Utilité normalisée

Le but de l'utilité normalisée (Hull et Robertson, 1999; Robertson et Hull, 2000) est d'atténuer l'effet des très grandes utilités absolues en normalisant l'utilité entre 0 et 1 ce qui permet de comparer plus significativement les systèmes de filtrage. Cette mesure d'utilité normalisée du système S pour le profil T est exprimée comme suit :

$$Utilité\ normalisée = \frac{max(u(S,T), U(n)) - U(n)}{MaxU(T) - U(n)} \tag{2.9}$$

avec :

$u(S,T)$: l'utilité absolue du système S pour le profil T (Eq. 2.8),

$U(n)$: l'utilité quand n documents non pertinents sont sélectionnés,

$MaxU(T)$: utilité maximale possible pour le profil T.

2.3.2.2 Mesures d'évaluation orientées rang

Un système de filtrage d'information doit décider de la pertinence d'un document en calculant un score de similarité par rapport au profil. Si ce score est supérieur à un seuil donné, alors il est accepté. Par ailleurs, en plus de la décision binaire (accepté/rejeté), un système de filtrage peut attribuer un score de confiance proportionnel à la probabilité de la pertinence d'un document. Si le système détermine que le document a une très forte chance d'être pertinent, il lui affecte un score de confiance maximal (par exemple 1000), sinon si la probabilité de pertinence de ce document est plus faible, le score de confiance sera aussi plus faible.

En se basant sur les scores de confiance, les mesures de précision et de rappel à un rang k peuvent être appliquées en supposant que les documents ayant les meilleurs scores de confiance correspondent aux documents des premiers rangs.

Frank *et al.* (2012) ont proposé une mesure ***hF1*** pour mesurer la performance d'un système de filtrage de documents pertinents à un ensemble de topics. En effet, en supposant que les scores de confiance d'un système de filtrage varient entre 1 et 1000, une F_{mesure} moyenne est calculée pour chaque point de confiance. La mesure hF1 correspond donc à la valeur maximale de F_{mesure} obtenue parmi tous les points de confiance.

$$hF1 = Max_i(F_{mesure}@i) \tag{2.10}$$

2.3.2.3 Mesures d'évaluation orientées gain

Dans certains cas de filtrage, par exemple lors du filtrage de documents décrivant un événement très connu comme une catastrophe naturelle, le flux peut contenir un très grand nombre de documents pertinents. Dans ce cas là, une nouvelle problématique apparaît. Par souci de gain de

temps, l'utilisateur ne souhaite pas que lui soit retourné tous les documents pertinents du flux qui pourraient reporter de l'information redondante, mais préfère un petit nombre de documents contribuant chacun à des nouvelles informations pertinentes qui enrichissent sa connaissance sur le sujet considéré (représenté par le profil).

Les mesures d'évaluation classiques (décrites dans la sous-section 2.2.4) considèrent qu'un document donné est soit pertinent ou non pertinent ou bien pertinent avec un certain degré (*NDCG*). Cependant, ces mesures ne permettent pas de mesurer la gain effectif acquis par l'utilisateur par un document (ou une phrase) étant donné qu'il a consulté des anciens documents (ou phrases). Pour mieux évaluer les systèmes de filtrage d'information, Aslam *et al.* (2013) ont proposé deux mesures similaires à la précision et au rappel classiques, mais tenant compte du gain cumulé par l'utilisateur par chaque phrase lue.

Soient un profil donné P et un flux de documents F. Soit l'ensemble $N = \{n_1, n_2, ..., n_m\}$ composé de m informations pertinentes par rapport à P. Chaque information n_i est associée à un degré d'importance donné par $R(n_i)$ et à une date d'apparition $n_i.date$.

Soit un système S renvoyant un ensemble de phrases U à partir des documents du flux F. Chaque phrase $u_j \in U$ est un couple $(u_j.texte, u_j.date)$, tel que $u_j.texte$ représente son texte et $u_j.date$ indique sa date de sélection. Une phrase u_j peut reporter une ou plusieurs informations pertinentes ($\in N$). Lorsqu'une phrase u_j reporte une information pertinente n_i, on dit que u_j est associée à l'information pertinente n_i et on note cette association de la façon suivante : $n_i \leftrightarrow u_j$. La vérité terrain est composée de l'ensemble des informations pertinentes N ainsi que les phrases associées (issues du *pool*).

L'évaluation du système S consiste à mesurer le gain apporté par les phrases sélectionnées U en les comparant aux informations pertinentes N. Pour ce faire, deux mesures sont appliquées : le gain attendu (EG) et l'exhaustivité (C).

Le gain attendu par un système S, similaire à la notion de précision, est calculé par l'équation suivante :

$$EG(S) = \frac{\sum_{\{i \in [1,m] : M(n_i,S) \neq 0\}} gain(M(n_i, S), n_i)}{|U|} \tag{2.11}$$

L'exhaustivité, similaire à la notion du rappel, mesure la capacité du système à couvrir autant d'informations pertinentes que possible, elle est calculée comme suit :

$$C(S) = \frac{\sum_{\{i \in [1,m] : M(n_i,S) \neq 0\}} gain(M(n_i, S), n_i)}{\sum_{i \in [1,m]} R(n_i)} \tag{2.12}$$

Avec

- $|U|$ est le nombre phrases renvoyées par le système S.
- $M(n_i, S)$ est une fonction qui renvoie la première phrase sélectionnée u (détectée plut tôt) pouvant être associée à n_i. Le gain de cette phrase par rapport à n_i est donné par la fonction $gain(M(n_i, S), n_i)$ qui peut être calculée de deux manières :
  - en pénalisant le temps de latence entre la date de l'apparition de l'information n et la date de sélection de la phrase u (qui contient l'information n), dans ce cas $gain(u, n) = R(n) * Latency(u.date, n.date)$, telle que $Latency(d_1, d_2)$ est une fonction qui pénalise le temps de latence entre les deux dates données en paramètres d_1 et d_2.
  - sans pénaliser le temps de latence, dans ce cas $gain(u, n_i) = R(n_i)$.

Illustrons par un exemple. Supposons qu'un système S renvoye 5 phrases $U = \{u_1, u_2, u_3, u_4, u_5\}$ telles que $u_1.date < u_2.date < u_3.date < u_4.date < u_5.date$. Soit $N = \{n_1, n_2, n_3, n_4\}$ l'ensemble des informations pertinentes à détecter ayant les degrés d'importance $R(n_1) = R(n_2) = 1$, $R(n_3) = 0.5$ et $R(n_4) = 0.75$. Soit la vérité terrain composée par les associations suivantes : $n_1 \leftrightarrow u_1$, $n_1 \leftrightarrow u_2$, $n_2 \leftrightarrow u_3$ et $n_3 \leftrightarrow u_3$. Autrement dit, l'information pertinente n_1 est détectée par les phrases $u1$ et $u2$; la phrase $u3$ reporte deux informations pertinentes n_2 et n_3 ; et l'information n_4 n'a pas été détectée par le système S. Les mesures $EG(S)$ et $C(S)$ sont calculées de la façon suivante :

$$\begin{aligned} EG(S) &= \frac{gain(M(n_1, S), n_1) + gain(M(n_2, S), n_2) + gain(M(n_3, S), n_3)}{|U|} \\ &= \frac{gain(u_1, n_1) + gain(M(u_3, S), n_2) + gain(M(u_3, S), n_3)}{5} \\ &= \frac{1 + 1 + 0.5}{5} = 0.500 \end{aligned}$$

$$\begin{aligned} C(S) &= \frac{gain(M(n_1, S), n_1) + gain(M(n_2, S), n_2) + gain(M(n_3, S), n_3)}{R(n_1) + R(n_2) + R(n_3) + R(n_4)} \\ &= \frac{gain(u_1, n_1) + gain(M(u_3, S), n_2) + gain(M(u_3, S), n_3)}{R(n_1) + R(n_2) + R(n_3) + R(n_4)} \\ &= \frac{1 + 1 + 0.5}{1 + 1 + 0.5 + 0.75} = 0.769 \end{aligned}$$

Nous voyons bien dans cet exemple que la mesure $EG(S)$ pénalise la redondance. En effet, le calcul du gain est de l'information n_1 est basé uniquement sur la phrase u_1, même si cette information a été également reportée dans la phrase u_2.

2.4 Conclusion

Dans ce chapitre, nous avons introduit les concepts de base de la recherche et du filtrage d'information. Ces deux types d'accès s'appuient sur des modèles de recherche afin d'évaluer la pertinence des documents vis-à-vis des requêtes/profils. Les modèles existant dans la recherche d'information sont arrivés à une certaine maturité afin de trouver les documents pertinents par rapport à une requête sous forme de mots clés. Cependant, de nouvelles problématiques sont apparues lorsque l'on considère des requêtes relatives à des entités (par exemple des personnes, des organisations, des lieux, des produits, etc.). Plusieurs travaux récents visent à aller au delà la représentation classique mots-clés versus documents en essayant de modéliser d'une manière explicite la notion d'entité. Nous détaillons dans le chapitre suivant un état de l'art sur les différentes tâches et modèles orientés entité.

Chapitre 3

Recherche d'information orientée entités

3.1 Introduction

Selon des analyses effectuées par Guo *et al.* (2009), environ 71% de requêtes utilisateurs contiennent des entités nommées. Comme nous le détaillerons dans la section 3.2, une entité est une "chose" qui peut être distinctement identifiée. Sans prendre en compte les spécificités des entités, un système de recherche ou de filtrage d'information classique peut ne pas être très performant pour traiter certaines tâches de recherche orientées entités telles que la *recherche d'entités ad-hoc* (Pound *et al.*, 2010), le *filtrage de documents centrés sur une entité* (Zhou et Chang, 2013; Frank *et al.*, 2013, 2014), le *résumé automatique centré sur une entité* (Aslam *et al.*, 2013, 2014) et la *recherche d'entités reliées* (Balog *et al.*, 2009, 2010; Vydiswaran *et al.*, 2009).

Dans ce chapitre, nous rappelons tout d'abord la définition de la notion d'entité (section 3.2), ainsi que les différentes sources d'informations sur des entités (section 3.3). Nous présentons ensuite, dans la section 3.4, les différentes tâches de recherche d'information orientée entités.

3.2 Notion d'entité

Littéralement, une entité est quelque chose qui existe en soi, réellement ou potentiellement, de façon concrète ou abstraite, physiquement ou pas. En informatique, le terme *entité* a été défini pour la première fois dans le monde des bases de données en 1976 par Chen comme étant une "chose" qui peut être distinctement identifiée (Chen, 1976). Une personne, une entreprise ou un événement sont des exemples d'entités.

Sur le web, Nie *et al.* (2012) définissent une *entité web* comme la principale unité de donnée à propos de laquelle des informations web doivent être collectées, indexées et classées. Il considère que les entités web sont généralement des concepts reconnaissables qui ont un intérêt pour un domaine d'application tels que des personnes, des organisations, des lieux, des produits, des articles scientifiques, des conférences ou des journaux.

Balog *et al.* (2009, 2010) définissent une *entité web* comme étant une "chose" typée contribuant à un résultat de recherche et identifiée par au moins une page internet (*homepage*) consacrée uniquement à elle. Par exemple, l'Institut de Recherche en Informatique de Toulouse (IRIT) est une entité web identifiée par sa page web *http ://www.irit.fr* et pouvant faire l'objet d'un résultat de recherche répondant à la requête "Quels sont les laboratoires de recherche à Toulouse qui s'intéressent à la *Recherche d'Information* ?". Cette définition met l'hypothèse que toute entité d'intérêt aurait au moins une page internet qui l'identifie.

Meij *et al.* (2014) ont proposé un tutoriel dans lequel une entité est définie comme étant une "chose" ou un objet unique, identifiable et ayant les propriétés suivantes

- un identifiant (*URI*),
- un ou plusieurs noms (appelés dans la littérature les variantes de l'entité),
- un ou plusieurs types,
- des attributs,
- des relations avec les autres entités.

Dans ce travail, nous utilisons la définition d'entité donnée par Meij *et al.* (2014). La figure 3.2 montre un exemple d'une entité web décrite par une page Wikipedia (une source d'information universelle décrivant des entités web et leurs propriétés). Nous pouvons voir toutes les différentes propriétés décrivant l'entité : l'identifiant *https ://fr.wikipedia.org/wiki/Andy_Murray*, le nom et les différentes variantes de l'entité (*Andy Murray, Andrew Murray*), le type (*joueur professionnel de tennis*), les attributs (*taille : 1.9m, poids : 84 kg, date de naissance : 15 mai 1987*, ...) et les relations avec les autres entités (*Entraîneurs : Amélie Mauresmo et Jonas Björkman, Naissance : Glasgow,* ...).

Les entités du monde réel peuvent évoluer au cours du temps : une personne peut effectuer des nouvelles actions. Un joueur de tennis peut gagner de nouveaux tournois, améliorer son classement mondial, etc. Une entreprise peut changer de situation, faire de nouvelles coopérations, etc. Ces évolutions sont reportées dans le Web via des articles d'actualités (*news*), commentées par des utilisateurs dans les réseaux sociaux ou encore illustrées

FIGURE 3.1 – Extrait de la page Wikipedia de l'entité *Andry Murray* (*01 Septembre 2015*)

dans les encyclopédies et les bases de connaissances qui constituent des sources d'informations permettant de décrire les entités de manière plus structurée et organisée.

3.3 Sources d'informations sur des entités

Les besoins en information des utilisateurs sont de plus en plus centrés sur des entités. Cet intérêt croissant a encouragé l'apparition de plusieurs projets décrivant les informations sur les entités de manière semi-structurée comme dans l'encyclopédie Wikipedia ou bien de manière structurée dans les bases de connaissances telles que Freebase, DBpedia, etc. Nous décrivons ci-dessous les principales caractéristiques de ces projets :

- ***Wikipedia*** : Wikipedia est une encyclopédie universelle généraliste

lancée en 2001 et devenue l'un des sites web les plus populaires dans le monde. Son contenu est alimenté de manière collaborative par plus de cent mille contributeurs à travers le monde. Elle comporte plus de 37 millions articles dans 288 langues [1]. La version anglaise est la plus riche avec 4.9 millions d'articles décrivant chacun une entité ou un concept. Généralement, le texte de l'article commence par une définition, un résumé ou une brève description de l'entité. Souvent, un tableau à côté de la synthèse propose des informations structurées sur l'entité appelées les *infoboxes*. Ces *infoboxes* contiennent les faits les plus importants sur l'entité décrite et sont affichées sous forme de paires attribut-valeur (Voir la figure 3.2).

- ***Linked Open Data (LOD)*** : Le LOD [2] est un projet collaboratif visant à publier des jeux de données ouverts et liés sur le Web en respectant les principes du Web sémantique édictés par Tim Berners-Lee (Bizer *et al.*, 2009). Le *LOD* comporte aujourd'hui plus de 570 jeux de données liés entre eux et répartis en 9 catégories : *Cross-Domain*, *Publications*, *Geographic*, *Media*, *Gouvernment*, *Life Sciences*, *User-generated Content Social Networking* et *Linguistics*. Ces jeux de données sont structurés sous forme de *triplets RDF* (Resource Description Framework) et peuvent être exploités non seulement par des humains, mais aussi par des programmes et agents logiciels. La figure 3.3 montre un extrait du nuage des jeux de données dans le *LOD*.

- ***DBpedia*** : DBpedia [3] est un projet universitaire et communautaire d'exploration et extraction automatiques de données à partir des infoboxes Wikipedia (Lehmann *et al.*, 2015). Les données extraites sont représentées en RDF sur la base d'un vocabulaire définis notamment par l'ontologie de DBpedia. Aujourd'hui, DBpedia occupe le centre du nuage du projet *Linked Open Data* (Figure 3.3).

- ***Freebase*** : Freebase [4] est une large base de connaissances généraliste développée par la société américaine de logiciels *Metaweb*. Elle a été lancée publiquement en Mars 2007. Ses données sont créées et maintenues d'une manière collaborative (Bollacker *et al.*, 2008). Freebase [5] compte plus de 39 millions de sujets (*topic*) sur les entités du monde réel. Les données sur ces entités sont structurées selon un schéma exprimé à travers des domaines (espaces de noms communs,

1. https ://en.wikipedia.org/wiki/Wikipedia :Size_of_Wikipedia
2. http ://linkeddata.org/
3. http ://wiki.dbpedia.org/about
4. http ://www.freebase.com/
5. https ://developers.google.com/freebase/guide/basic_concepts

par exemple, le domaine *Sports*), des types et des propriétés. Une entité dans Freebase peut avoir plusieurs types. Par exemple, l'entité `Leonardo DiCaprio` est associée à plusieurs types comme `personne`, `acteur de film`, `acteur de TV`, etc. Un type particulier est décrit par un ensemble de propriétés. Par exemple, le type personne est caractérisé par une `date de naissance`, un `lieu de naissance`, un `genre`, etc. alors que le type acteur de film est caractérisé par la propriété `Performances de film`. Les propriétés permettent d'une part de décrire les entités et d'autre part de les lier entre elles formant ainsi un graphe d'entités.

- ***Le Google Knowledge Graph*** : Le *Google Knowledge Graph*[6] est une base de connaissances utilisée par le moteur de recherche de Google, depuis Mai 2012, pour compiler les résultats de recherche avec des informations sémantiques issues de sources diverses telles que *CIA World Factbook*, Freebase et Wikipédia (Singhal, 2012). Les entités sont décrites et reliées par un graphe. Cependant, le format de données utilisé n'est pas connu.

Remarque : Dans la suite de ce mémoire, nous appelons ces sources des **bases de connaissances**, bien que les travaux d'ingénierie considèrent que Wikipedia est une encyclopédie et non pas une base de connaissances car la connaissance n'est pas formellement décrite.

3.4 Recherche d'information orientée entités

Dans cette section, nous présentons les tâches de recherche d'information orientée entités indiquées en gras dans la figure 3.3. Dans la tâche de *recherche d'entités ad-hoc* (Pound *et al.*, 2010), les entités représentent la sortie du système de recherche, alors que dans les tâches de *filtrage de documents centrés sur une entité* (Zhou et Chang, 2013; Frank *et al.*, 2013, 2014) ou de *résumé automatique centré sur une entité* (Aslam *et al.*, 2013, 2014), les entités sont fournies en entrée du système. Pour la tâche de *recherche d'entités reliées* (Balog *et al.*, 2009, 2010; Vydiswaran *et al.*, 2009), une entité et un narratif sont donnés en entrée, les résultats attendus sont les entités reliées à l'entité donnée par la relation décrite dans le narratif.

Dans la suite, nous distinguons deux axes de recherche d'information orientée entités :

- Dans le *premier axe*, l'utilisateur cherche une liste d'entités en réponse à sa requête et non pas une liste classique de documents. Nous présentons dans la section 3.4.1 les différentes applications et

6. https ://www.google.com/intl/fr_fr/insidesearch/features/search/knowledge.html

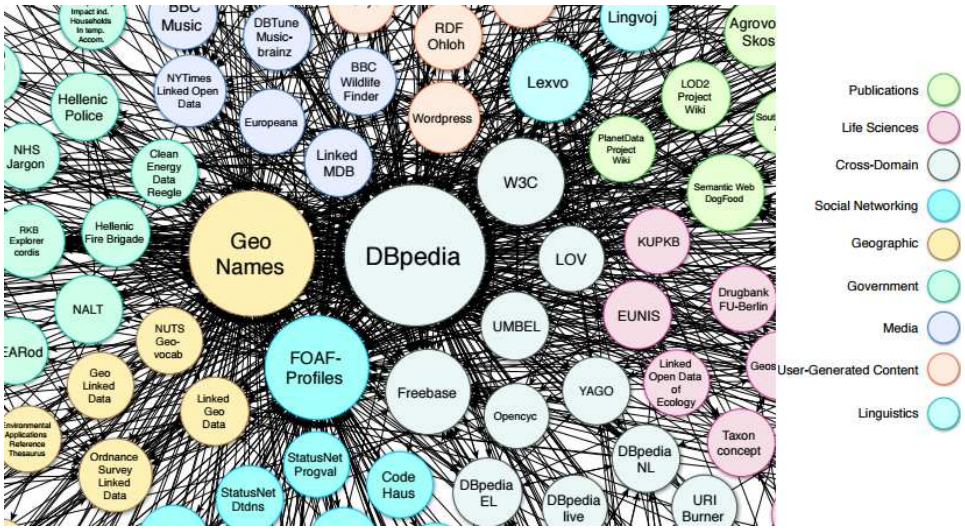

FIGURE 3.2 – Extrait du nuage de jeux de données du projet *Linked Open Data* (Août 2014)

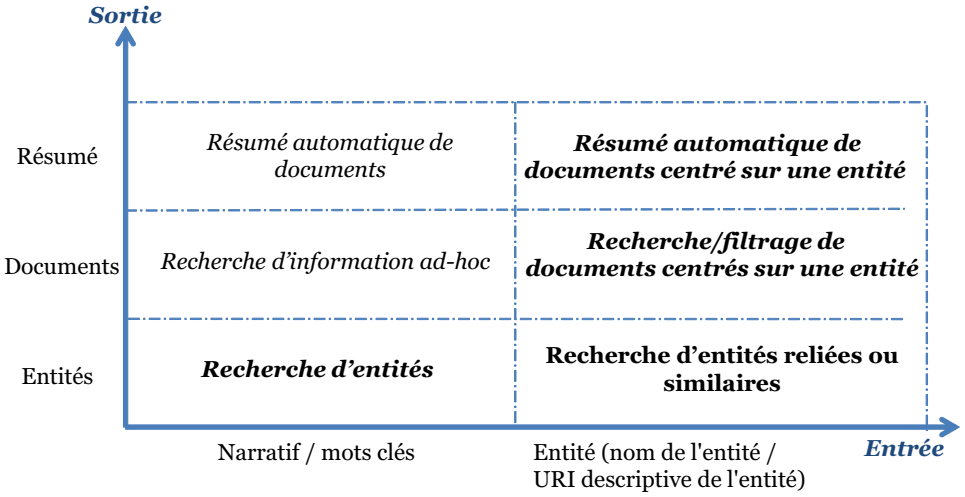

FIGURE 3.3 – Quelques scénarios de recherche d'information. Les opérations indiquées en gras concernent la recherche d'information orientée entités

campagnes d'évaluation en lien avec la recherche d'entités.

- Dans le *deuxième axe*, l'utilisateur s'intéresse à des informations sur une entité donnée en entrée. Les informations filtrées peuvent être présentées sous forme de :
  - **documents**. Nous appelons ce paradigme ***filtrage de documents centrés sur une entité*** que nous détaillons dans la section 3.4.2.
  - **un résumé** agrégeant les informations importantes, plus précisément les phrases contenant des informations importantes. Ce paradigme est appelé ***résumé temporel centré sur une entité*** que nous détaillons dans la section 3.4.3.

  Les travaux de cette thèse se focalisent sur ce *deuxième axe* que nous nommons ***filtrage et agrégation d'information autour d'une entité***.

3.4.1 Recherche d'entités

Chercher des entités sur le Web est une nouvelle problématique de recherche qui va au-delà de la recherche classique de documents. Récemment, il a été observé que les entités (plutôt qu'une liste classique de documents) répondent mieux à certaines requêtes du Web (Balog *et al.*, 2009, 2010). Par exemple, lorsqu'un utilisateur cherche à trouver "la liste des pays africains qui parlent l'anglais", un moteur de recherche d'information classique renvoie des documents concernant les pays africains et la langue anglaise. Cependant, c'est à l'utilisateur de parcourir les documents renvoyés afin d'extraire les noms des pays recherchés.

Une variétés de tâches de recherche d'entités existent (Balog *et al.*, 2009; Pound *et al.*, 2010; Kopliku *et al.*, 2014). Nous les présentons dans la sous-section suivante.

3.4.1.1 Taxonomie des tâches de recherche d'entités

- ***Recherche d'entités ad-hoc (Ad-hoc Entity Retrieval)*** : La recherche d'entités ad-hoc est une tâche qui s'intéresse à renvoyer des entités en réponse à une requête utilisateur cherchant à trouver non pas une liste de documents mais plutôt une liste d'entités de même type (spécifié dans la requête) (Pound *et al.*, 2010; Kopliku *et al.*, 2014). Les entités résultats ont le même type, mais un système de recherche d'entités ad-hoc doit être capable de répondre aux requêtes ciblant différents types d'entités. La figure 3.4 montre un exemple de recherche d'entités ad-hoc (universités situées à Toulouse).

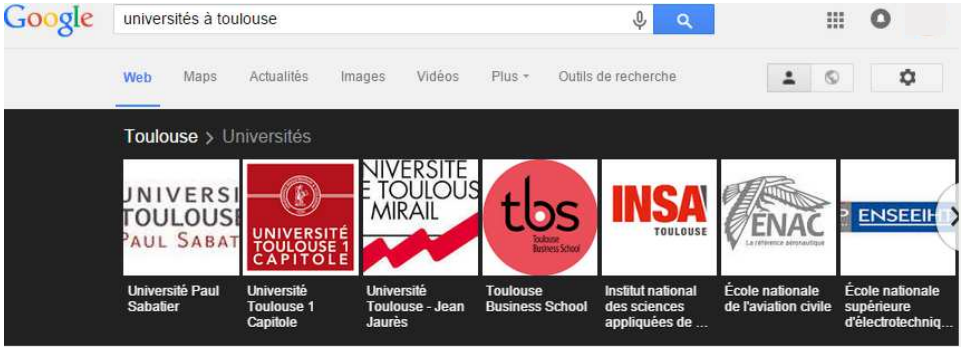

FIGURE 3.4 – Exemple de recherche d'entités ad-hoc par le moteur de recherche Google (requête soumise le 01/10/2015)

- ***Complétion de la liste d'entité (Entity List Completion)*** : C'est une tâche similaire à la tâche recherche d'entités à ceci près que l'utilisateur doit fournir, en plus de sa requête, des exemples d'entités pertinentes (Balog *et al.*, 2009, 2010; Kopliku *et al.*, 2010). Par exemple, pour la requête "Liste des régions françaises", l'utilisateur doit donner quelques exemples comme "`Île-de-France`", "`Alsace`", "`Midi-Pyrénées`". Le système devra renvoyer toutes les entités pertinentes qui ne sont pas fournies par l'utilisateur.

- ***Question/Réponse (Question Answering)*** : L'objectif de la tâche question-réponse est de promouvoir des systèmes qui retournent directement des réponses, plutôt que des documents contenant des réponses, en ne considérant que la requête posée par une question en langage naturel (Voorhees, 2004; Hirschman et Gaizauskas, 2001). Plusieurs catégories de questions peuvent être proposées comme les questions factuelles (Qui, Quand, Ou, ...), booléennes (oui/non), définition (Qu'est-ce que), complexes (Pourquoi, comment) ou des questions de type liste (par exemple, "Quels pays ont été touchés par le tremblement de terre Népal 2015 ?"). Dans la tâche de Question/Réponse, l'utilisateur attend une réponse précise par exemple un numéro, un nom de lieu ou une liste d'éléments (entités ou autres). Lorsque la question posée cible une liste d'entités, la tâche Question/Réponse devient similaire à la recherche d'entités ad-hoc.

- ***Recherche d'entités reliées (Related Entity Finding)*** : C'est une tâche qui cherche à renvoyer des entités qui s'engagent dans une relation donnée avec une entité source donnée (Balog *et al.*, 2009, 2010). La requête est décrite en langage naturel et doit mentionner une entité source, sa relation avec les entités attendues et leur type. Par exemple, un utilisateur souhaite savoir la liste des "compagnies

aériennes qui utilisent actuellement l'avion `Boeing-747`". Dans cet exemple, l'entité source est "`Boeing-747`" et sa relation avec les entités cibles est "utiliser".

- ***Recherche d'experts (Expert Finding)*** : Cette tâche est une variante de la tâche de recherche d'entités ad-hoc ciblée sur les entités de type personne. Elle s'intéresse particulièrement à rechercher des experts sur un sujet précis (Craswell *et al.*, 2001; Macdonald et Ounis, 2006). Dans le contexte de l'entreprise, cela consiste à trouver une personne (employé, un associé, ...) qui a une connaissance particulière sur un sujet donné (C++, Hadoop, ...). L'objectif pour l'entreprise est de faciliter le recrutement d'équipe ou d'une personne spécialisée sur un besoin. Dans le contexte des conférences, ce type de tâche permet d'aider les organisateurs à affecter les articles soumis aux membres du comité de lecture en fonction de leurs expertises et domaines intérêts.

- **Autres variations** : Plusieurs autres variations de recherche d'entités existent et s'intéressent à un type particulier d'entités comme la recherche de produits (Amazon Product Search [7], Google Product Search [8], Yahoo Shopping [9], ...), la recherche des publications bibliographiques (Microsoft Libra [10], Google Scholar [11]), la recherche de livres (Google Book Search) [12], etc.

3.4.1.2 Campagnes d'évaluation relatives à la recherche d'entités

Les campagnes d'évaluation TREC, INEX et SemSearch se sont intéressées à la recherche d'entités. Le tableau 3.1 présente quelques collections de test proposées pour l'évaluation de la recherche d'entités.

Certaines différences existent entre ces tâches au niveau de type d'entités traitées, le type des sources à interroger et le format des résultats.

- ***Type d'entités :*** La campagne *TREC Entreprise* a proposé la tâche *Expert Finding* qui s'intéresse uniquement à retrouver des entités de types personnes, alors que les autres campagnes se sont intéressées à renvoyer n'importe quel type d'entités.

- ***Type de sources :*** Les sources dans lesquelles les entités doivent retrouvées peuvent être :

7. http ://www.amazon.com/
8. http ://www.google.com/shopping
9. https ://shopping.yahoo.com/
10. http ://academic.research.microsoft.com/
11. https ://scholar.google.com
12. https ://books.google.com/

Campagne	Tâche	Sources et types	Résultats
TREC Entreprise (2005-2008)	Expert Finding	Intranet d'une organisation (non-structuré)	Liste de personnes (noms)
TREC Entity (2009-2011)	Related Entity Finding ; List Completion	ClueWeb09 (non-structuré)	Liste d'entités (homepage URI)
INEX Entity Ranking (2007-2009)	Entity Ranking ; Entity List Completion	Wikipedia Semi-structuré (XML)	Liste d'entités (Wikipedia URI)
SemSearch Challenge (2010-2011)	Entity Search ; List Search	BTC2009 structuré (RDF)	Liste d'objets (URI)
INEX Linked Data (2012-2013)	Ad-hoc search	DBpedia structuré	Liste d'objets (DBpedia URI)

TABLE 3.1 – Collections de test proposées pour l'évaluation de la recherche d'entités

  - **Non structurées** : Les campagnes *TREC Entreprise* et *TREC Entity* utilisent des documents non structurés. Dans la première campagne, les documents sont issus de l'Intranet des organisations W3C et CSIRO. Dans la deuxième campagne les documents sont issus du Web.
  - **Semi-structurées** : *INEX Entity Ranking* se base sur la collection XML de Wikipedia.
  - **Structurées** : Les campagnes *INEX Linked Data* et *SemSearch Challenge* utilisent des données structurées en RDF.

- ***Format des résultats :*** Dans la tâche *Expert Finding* les résultats sont sélectionnés à partir de la liste des personnels de l'entreprise, alors que dans les autres tâches les systèmes doivent retrouver l'*URI* de l'entité dans la collection.

3.4.2 Filtrage de documents centrés sur une entité

Comme nous l'avons évoqué dans la section 3.3, aujourd'hui, les bases de connaissances telles Wikipedia, DBpedia et Freebase représentent des sources principales pour accéder aux informations disponibles sur une grande variété d'entités. En exploitant ces sources, un utilisateur peut facilement retrouver les principales informations sur une entité donnée sans devoir faire l'effort de rechercher et parcourir les documents renvoyés par un moteur de recherche classique. Cependant, ces sources d'informations ne sont pas mises à jour de manière automatique et peuvent par conséquent ne pas contenir les informations récentes sur certaines entités.

Un utilisateur cherchant principalement les nouvelles informations sur un sujet donné peut se servir d'un système de filtrage d'information qui analyse un flux dynamique de documents à la différence d'un système de recherche d'information classique qui cherche l'information dans une collection statique.

Le filtrage de documents centrés sur une entité (figure 3.5) est un nouveau paradigme de recherche qui s'intéresse à renvoyer à partir d'un flux continu de documents ceux qui sont pertinents par rapport à une entité donnée décrite par un profil appelé *profil d'entité*. Ce paradigme peut être utilisé dans plusieurs cadres d'applications que nous décrivons dans la sous-section suivante.

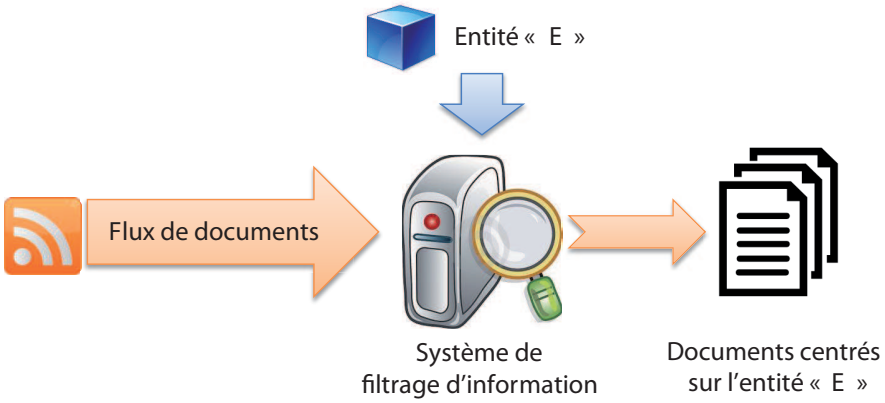

FIGURE 3.5 – Système de filtrage de documents centrés sur des entités

3.4.2.1 Applications du filtrage de documents centrés sur une entité

Le filtrage de documents centrés sur une entité donnée peut se retrouver dans plusieurs applications telles que :

- ***le suivi des célébrités*** (Zhou et Chang, 2013) : Beaucoup de personnes aiment suivre les activités de leurs célébrités préférées. Microsoft a développé récemment une nouvelle application mobile "*SNIPP3T*" [13] qui représente un exemple de système de filtrage d'information centrée sur les entités célébrités.

- ***l'informatique décisionnelle (en anglais, Business Intelligence)*** (Zhou et Chang, 2013) : Dans le cadre d'une

13. http ://microsoft-news.com/microsoft-updates-snipp3t-celebrity-news-app-with-brand-new-snip-feature/

entreprise, collecter automatiquement les avis des utilisateurs sur les entités produits est une tâche utile pour améliorer la qualité des futurs produits. En outre, dans une perspective de gestion de crise, surveiller les avis des utilisateurs sur le Web permet à l'entreprise de détecter les crises potentielles ou émergentes dans un temps court.

- ***le suivi des événements en temps réel*** (Aslam *et al.*, 2013, 2014; Xu *et al.*, 2013a) : Lors d'un événement de crise (comme une catastrophe naturelle, un attentat, ...), les utilisateurs souhaitent avoir rapidement les dernières informations sur cet événement, surtout s'ils sont directement touchés. Un système de filtrage d'information centré sur des entités événements permet de détecter les informations importantes sur ces événements au fur et à mesure qu'elles sont publiées dans le Web.

- ***la détection des nouvelles valeurs d'attributs des entités (en anglais, Streaming Slot Filling)*** (Frank *et al.*, 2014; Surdeanu et Heng, 2014) : Cette tâche consiste à analyser un flux de documents Web et détecter les changements des valeurs associées aux propriétés des entités (valeurs d'attributs, entités reliées). Par exemple, pour une entité de type personne, le but est de détecter les nouvelles valeurs associées aux attributs taille, age, date de décès ou les nouvelles entités reliées par des relations prédéfinies comme membreDe, fondateurDe. etc.

- ***la recommandation de citations cumulatives (en anglais, Cumulative Citation Recommendation - CCR)*** : La plupart des bases de connaissances sont maintenues manuellement par des éditeurs bénévoles. Aujourd'hui, la mise à jour de ces bases de connaissances devient de plus en plus difficile en raison du nombre limité de contributeurs et de l'énorme volume d'entités. Frank *et al.* (2012) ont indiqué que pour un ensemble d'entités non populaires, le temps de latence médian entre la date à laquelle une page Wikipedia est enrichie par une citation d'un document du Web et la date à laquelle ce document est publié sur le Web est de 365 jours.
  La tâche *Cumulative Citation Recommendation (CCR)* (Frank *et al.*, 2012, 2013, 2014) a été introduite dans la campagne d'évaluation TREC dans le but d'accélérer la mise à jour des bases de connaissances. Elle consiste à analyser un flux de documents Web et renvoyer ceux qui sont pertinents par rapport aux entités d'une base de connaissances comme Wikipedia. Un document est considéré comme pertinent par rapport à une entité donnée s'il contient des informations pertinentes sur cette entité et pourra donc être cité comme référence dans la page descriptive de l'entité. Les documents pertinents sont détectés dès

qu'ils sont publiés, et recommandés aux contributeurs de la base de connaissances.

Dans la tâche CCR, deux niveaux de pertinence peuvent être distingués :
- ***document vital*** (*vital document*) : c'est un document qui contient une ou plusieurs informations pertinentes et opportunes sur l'entité comme par exemple un document décrivant une nouvelle action ou situation concernant l'entité. La détection d'un document vital implique la mise à jour immédiate de la page descriptive de l'entité.
- ***document utile*** (*useful document*) : c'est un document qui contient des informations pertinentes mais connues sur l'entité.

Les documents non pertinents peuvent être classés à leur tour en deux classes :
- Neutre (*neutral*) : Un document est considéré comme neutre s'il mentionne une information non importante sur l'entité qui ne peut pas être citée dans sa page descriptive,
- Déchet (*garbage*) quand le document ne fournit aucune information sur l'entité. Un document qui mentionne le même nom que l'entité mais qui parle d'une autre entité est considérée comme *Garbage*.

3.4.2.2 Critères de pertinence

Nous présentons dans cette section des différents critères exploités dans les travaux de l'état de l'art pour filtrer les documents pertinents par rapport à des entités d'intérêt.

3.4.2.2.1 Critères de pertinence a priori

Filtrer les documents centrés sur une entité donnée nécessite l'établissement d'une correspondance entre deux éléments : l'entité et le document à analyser. Cependant, considérer des caractéristiques a priori sur chacun de ces éléments peut améliorer la performance du filtrage. Dans ce cadre, deux catégories de critères a priori ont été proposées.

La première catégorie de critères de pertinence a priori s'intéresse aux propriétés d'un document indépendamment de l'entité. Balog *et al.* (2013); Balog et Ramampiaro (2013) ont exploité la longueur, la source (social, news, ...) et la langue du document comme critères a priori caractérisant le document. Wang *et al.* (2013) ont utilisé le jour correspondant à date de publication du document comme un critère de plus qui peut caractériser le document. Bouvier et Bellot (2013) ont exploité un critère booléen indiquant si le document contient un titre ou non. Ils ont utilisé aussi

l'entropie de Shannon pour mesurer la quantité d'information délivrée par le document. D'autres critères plus avancés peuvent être exploités comme la crédibilité (Weerkamp et de Rijke, 2008) et la lisibilité (Polajnar *et al.*, 2012) du document.

La deuxième catégorie de critères de pertinence a priori vise à décrire les propriétés de l'entité indépendamment des documents de flux. Plusieurs travaux ont exploité le nombre d'entités reliées récupérées à partir de la page descriptive de l'entité (Balog *et al.*, 2013; Wang *et al.*, 2013; Liu *et al.*, 2013b; Bouvier et Bellot, 2013). Wang *et al.* (2015b) ainsi que Araújo *et al.* (2012) ont également exploité les types et les catégories auxquelles l'entité appartient. Ces informations sont collectées à partir de la page Wikipedia de l'entité ou bien en interrogeant DBpedia à l'aide d'une requête SPARQL cherchant à retrouver les valeurs associées aux propriétés *type* et *subject*.

3.4.2.2.2 Critères de pertinence entité-document

Les critères de pertinence entité-document visent à établir une correspondance entre l'entité et le document. Ils peuvent être vus comme des fonctions à deux paramètres : le document et l'entité.

Les critères entité-document les plus exploités concernent le nombre d'occurrences de l'entité dans le document, les positions des mentions de l'entité dans le document, la propagation des mentions de l'entité dans le corps du document calculée par la distance entre la première et la dernière mention de l'entité dans le document. Ces critères ont été exploités dans plusieurs travaux (Balog *et al.*, 2013; Wang *et al.*, 2013, 2015b; Bouvier et Bellot, 2013; Abbes *et al.*, 2013) avec certaines variantes pour le calcul des mesures qui peut se faire de manière absolue ou normalisée par rapport à la longueur du document permettant ainsi d'obtenir des valeurs entre 0 et 1. De plus, ils peuvent être calculés en considérant soit le nom complet de l'entité ou bien son nom partiel (par exemple, le prénom d'une personne).

Le comptage du nombre de mentions des entités reliées dans le document peut aussi être considéré comme un critère de pertinence. Liu *et al.* (2013b) ont proposé de pondérer les entités reliées et de les exploiter dans l'attribution des scores de documents.

En plus de l'exploitation des mentions des entités, il est utile d'exploiter les informations contenues dans les pages descriptives de l'entité. Plusieurs critères basés sur le calcul de la similarité ou de la divergence entre le document et la page descriptive de l'entité peuvent être exploités (Balog *et al.*, 2013; Wang *et al.*, 2013, 2015a). Les mesures les plus exploitées sont le cosinus, la similarité de Jaccard, l'inverse de la KL-divergence (Kullback et Leibler, 1951).

La page descriptive de l'entité peut citer des documents externes. Wang *et al.* (2013) pensent que ces citations sont extrêmement précieuses dans l'identification des documents pertinents. Ils ont proposé des critères de citations mesurant la similarité (Cosinus et Jaccard) entre le document et chacun des documents cités. Si la page descriptive de l'entité ne présente pas de citations, une alternative est d'utiliser le nom de l'entité comme requête à soumettre à un moteur de recherche et ensuite supposer que les premiers documents renvoyés comme des citations (Wang *et al.*, 2015b).

3.4.2.2.3 Critères de pertinence temporels

Les critères de pertinence entité-document ne sont pas capables de représenter les caractéristiques dynamiques des entités qui peuvent évoluer au cours du temps. Il a été souligné dans plusieurs travaux (Alonso *et al.*, 2011; Campos *et al.*, 2014; Moulahi *et al.*, 2015) que la prise en compte de l'aspect temporel permet d'améliorer la performance des méthodes de recherche d'information. Concernant le filtrage d'information autour d'entités, plusieurs critères temporels ont été proposés. Ces critères visent à capturer si quelque chose se produit autour de l'entité dans un laps de temps particulier en détectant s'il y a une augmentation brusque ou anormale d'un intérêt sur l'entité appelée rafale (*burst*) (Zhu et Shasha, 2003).

La détection des rafales autour de l'entité peut se faire de manière interne en détectant un changement dans le volume de documents mentionnant l'entité dans le flux (Bouvier et Bellot, 2013; Abbes *et al.*, 2013; Balog *et al.*, 2013; Wang *et al.*, 2013, 2015b). Plus le nombre de documents mentionnant l'entité à l'instant t est grand, plus la probabilité de considérer l'entité comme "explosive" est élevée.

Il est possible aussi d'exploiter des ressources externes pour détecter les rafales autour de l'entité. Balog *et al.* (2013); Wang *et al.* (2013, 2015b) ont exploité l'évolution du nombre de vues de la page Wikipedia de l'entité au moment ou quelques heures avant l'apparition du document. Pour les entités n'ayant pas une page Wikipedia, l'outil Google Trends [14], qui permet de fournir l'historique du volume de recherche normalisé de l'entité pour chaque jour, peut être exploité pour détecter les rafales (Wang *et al.*, 2015b).

Un ensemble de méthodes et d'algorithmes ont été proposés pour détecter et mesurer la force des rafales dans un flux de documents. Kleinberg (2002) a modélisé le flux de documents par un automate à un nombre infini d'états dans lequel les rafales apparaissent naturellement comme des transitions d'états. Cette méthode a été adoptée par He *et al.* (2007); Wang *et al.*

14. http ://www.google.com/trends/

(2015b) en se limitant à un modèle d'automates finis à deux états permettant d'identifier des éléments situés dans des rafales. Zhu et Shasha (2003) ont proposé une structure de données générale pour la détection efficace des rafales. Vlachos *et al.* (2004) ont proposé une méthode de détection des rafales en utilisant les meilleurs coefficients dans une transformée de Fourier.

3.4.2.2.4 Critères de pertinence basés sur des modèles de patrons

L'idée derrière ces critères est qu'un document pertinent mentionne certains modèles de patrons (*patterns*). Araújo *et al.* (2012) ont proposé une approche basée sur l'apprentissage des préfixes et des suffixes d'une entité à partir d'un ensemble de documents d'apprentissage annotés comme pertinents ou non pertinents. Les modèles de patrons gardés sont ceux qui apparaissent uniquement dans les documents pertinents et seront utilisés pour détecter des documents récemment apparus dans le flux.

Jiang *et al.* (2014) ont observé que les documents vitaux (pertinents et apportant de nouvelles informations) décrivent généralement l'entité en train d'effectuer des actions comme gagner un prix, lancer un nouveau album, organiser un événement, etc. Ils ont créé des modèles de patrons d'actions à partir d'un ensemble de documents échantillons vitaux. Ces modèles de patrons sont sous forme de *"entité + verbe normalisé"* si l'entité représente le sujet de la phrase, ou bien sous forme de *"verbe normalisé + entité"* si l'entité est l'objet. Par exemple, à partir de la phrase "Obama Won the Election", le modèle de patron identifié est "Obama win". Les auteurs ont appliqué un critère binaire indiquant si le document analysé présente ou pas un patron d'action.

3.4.2.3 Approches de filtrage de documents centrés sur une entité

Plusieurs approches ont été proposées pour filtrer les documents centrés sur une entité. Ces travaux exploitent un ou certains critères de pertinence décrits dans la section précédente. Nous pouvons distinguer trois catégories d'approches :

- ***Approches de classification***
  La première catégorie d'approches a traité la tâche en tant qu'un problème de classification (Balog *et al.*, 2013; Bonnefoy *et al.*, 2013; Wang *et al.*, 2013, 2014; Jiang *et al.*, 2014). Le but est de prédire pour chaque document sa classe (vital, utile, neutre ou déchet dans le cadre de la tâche CCR). Cette prédiction peut se faire en une seule étape avec un seul classifieur ou bien en plusieurs étapes en combinant plusieurs classifieurs en cascade.

Kenter (2013) a utilisé une classification bayésienne naïve et multinomiale qui se base sur la présence ou l'absence des termes du vocabulaire comme critères binaires. Bonnefoy *et al.* (2013) ainsi que Balog *et al.* (2013) ont proposé une approche de classification basée sur une ou plusieurs étapes pour apprendre la pertinence d'un document en se focalisant sur quatre familles de critères (1) *critères liés au document* comme la longueur et la source du document ; (2) *critères liés à l'entité* comme le nombre d'entités reliées dans DBpedia ; (3) *critères liés au couple document-entité* qui décrivent la relation entre le document et l'entité comme le nombre d'occurrence des mentions de l'entité dans le document et (4) *critères temporels* essayant de capturer si quelque chose se produit autour de l'entité à un point donné dans le temps en analysant les changements dans le volume de flux et du nombre de vues de la page Wikipedia de l'entité.

Wang *et al.* (2013) ont utilisé un classifieur Random Forest avec les mêmes familles de critères que celles utilisées dans (Balog *et al.*, 2013) et (Bonnefoy *et al.*, 2013) en considérant en plus des *critères de citations* reflétant la similarité entre un nouveau document et les documents cités dans la page Wikipedia de l'entité. Dans un travail ultérieur, Wang *et al.* (2015b, 2014) se sont intéressés de nouveau aux critères temporels pour la capture des rafales. Étant donné que les entités peuvent ne pas avoir de pages Wikipedia, ils ont proposé de détecter les périodes de *rafales* en exploitant les statistiques de *Google Trends*.

Jiang *et al.* (2014) ont remarqué que plusieurs critères utilisés dans les travaux précédents ne sont pas d'une grande nécessité. Ils ont proposé une approche de classification en gardant uniquement les critères temporels. Ils ont proposé par ailleurs trois nouveaux critères qui ont pour but de mieux distinguer les documents vitaux des documents utiles. Le premier critère est binaire indiquant si le document contient ou non un patron d'action composé de l'entité et d'un verbe d'action (par exemple *Peter Goldmark win*). Le deuxième critère se base sur l'extraction de certaines informations spécifiques sur l'entité comme le titre, la profession et l'adresse afin de construire un profil d'entité local qui sert à désambiguïser l'entité. Le troisième critère consiste à créer des groupes de documents similaires (clusters) qui sont enrichis dans le temps et à favoriser les premiers documents de chaque cluster car ces documents ont tendance à apporter des informations nouvelles sur l'entité.

- ***Approches de classement (ranking)***
  La deuxième catégorie d'approches a abordé la tâche comme un

problème de classement. L'idée est d'attribuer un score de pertinence à chaque document. Les documents ayant un score supérieur à un seuil donné seront sélectionnés. Dietz et Dalton (2013) ont utilisé une technique d'expansion de requête par des noms d'entités reliées pour filtrer les documents centrés sur l'entité d'intérêt. La méthode proposée par Liu *et al.* (2013b) classe les documents en exploitant le nombre d'occurrences et le poids des entités reliées récupérées en analysant la page Wikipedia de l'entité d'intérêt.
Ces travaux sont performants pour filtrer les documents pertinents sur une entité. Cependant, ils ne tentent pas de faire la distinction entre les documents vitaux et les documents utiles. Une raison possible est que ces modèles tentent de capter la topicalité (le sujet) plutôt que des caractéristiques temporelles.

Les approches de classification décrites dans le paragraphe précédent peuvent aussi être appliquées comme des méthodes d'apprentissage de classement (*Learning to Rank*) en considérant l'échelle de pertinence suivant (vital> utile> neutre> déchet) et en exploitant les mêmes critères (Wang *et al.*, 2014; Balog et Ramampiaro, 2013; Balog *et al.*, 2013). Selon des analyses effectuées dans (Balog et Ramampiaro, 2013) et (Wang *et al.*, 2014), les approches d'apprentissage de classement sont plus adaptées pour la tâche de filtrage de documents centrés sur une entité.

- ***Approches booléennes***
  Les deux premières catégories d'approches attribuent un score à chaque document basé sur la probabilité d'appartenir à une classe pertinente, ou bien sur une fonction de pertinence. Efron *et al.* (2014) a proposé un troisième type d'approche basée sur l'apprentissage de requêtes booléennes apprises à partir d'un ensemble de documents échantillons. Ces requêtes sont appliquées de façon déterministe pour filtrer les documents pertinents à l'entité.

Discussion sur les approches de l'état de l'art et leurs limites
La plupart des approches de l'état de l'art peuvent se classer selon deux *stratégies* :

- La première stratégie consiste à classifier ou faire un apprentissage de classement (*Learning To Rank*) des documents en se basant sur des annotations établies durant la période d'apprentissage. Cette stratégie se base sur la combinaison d'un grand nombre de critères (plus de 60) dont l'importance peut varier considérablement. Cette stratégie est néanmoins peu explicite puisqu'elle se base sur un (ou plusieurs) classifieur qui peut être vu comme une boîte noire.

- La seconde stratégie consiste à calculer un score de pertinence pour chaque document. Le document sera accepté si son score est supérieur à un seuil donné. Cette stratégie exploite un nombre limité de critères (facteurs) visant à calculer un score de pertinence pour chaque document. Les approches proposées utilisant cette stratégie (Liu *et al.*, 2013b; Dietz et Dalton, 2013) s'intéressent particulièrement à la "topicalité" (pertinence thématique) et négligent l'aspect temporel qui est très important dans une tâche de filtrage.

Un autre point important pour un système de filtrage de documents centrés sur des entités est sa capacité à traiter n'importe quelle entité. Certaines approches proposées sont **supervisées**, c'est-à-dire qu'elles nécessitent des documents d'entraînement pour apprendre un modèle pour chaque entité, ce modèle étant ensuite appliqué pendant la période d'évaluation. Ce type d'approche est très performant puisqu'il apprend les caractéristiques propres à chaque entité (Wang *et al.*, 2014). Cependant, il n'est pas raisonnable de l'appliquer lors du passage à l'échelle.

Plusieurs approches **semi-supervisées** se sont intéressées à résoudre ce problème en essayant d'apprendre un modèle générique qui peut s'appliquer à n'importe quelle entité (Balog *et al.*, 2013; Bonnefoy *et al.*, 2013; Wang *et al.*, 2013; Bouvier et Bellot, 2013). Néanmoins, les performances sont moins bonnes par rapport aux approches supervisées car les caractéristiques des différentes entités sont perdues dans ces méthodes.

Zhou et Chang (2013) ont proposé une méthode de transfert d'apprentissage qui exploite des critères associant des mots-clés entre les entités d'apprentissages et les entités de test. En effet, chaque critère est une paire <mot-clé w, entité e>. Chaque critère est décrit lui aussi par un vecteur de critères dit "méta-critères" $f(w, e)$ qui caractérisent l'importance potentielle d'un mot-clé w pour l'entité e. Par exemple, le méta-critère $IdPagePos(w, e)$ représente la première position du mot w dans la page descriptive de l'entité e (par exemple sa page Wikipedia), et dans la plupart du temps, les mots clés mentionnés plus haut dans cette page ont tendance à être plus importants. Un deuxième exemple de méta-critère représente la fréquence du mot-clé w dans l'infoBox de la page Wikipedia de l'entité. Si cette fréquence est différente de zéro, le mot-clé w aura plus tendance à être important.

Par exemple, considérons que `Mark Zuckerberg` comme une entité d'apprentissage et `Larry Page` comme une entité de test. Étant donné que l'exemple d'apprentissage montre que le mot "Facebook" est important pour l'entité `Mark Zuckerberg` et en associant le mot "Facebook" à "Google" car $f(Facebook,\ Mark\ Zuckerberg) \approx f(Google,\ Larry\ Page)$, le modèle pourra inférer que le mot "Google" est important aussi pour l'entité `Larry Page`.

Des travaux récents ont proposé d'apprendre des modèles de classification par classe d'entités qui semblent partager des caractéristiques communes (Wang *et al.*, 2015b; Bouvier et Bellot, 2015). Cette idée de regroupement est intéressante pour assurer un fonctionnement des systèmes de filtrage avec un taux de supervision faible tout en gardant au maximum des caractéristiques des entités.

3.4.2.4 Évaluation du filtrage de documents centrés sur une entité

Dans le contexte de filtrage de documents centrés sur une entité, la campagne d'évaluation TREC a lancé une nouvelle piste nommée *Knowledge Base Acceleration* qui s'est déroulée en 2012, 2013 et 2014. La tâche principale dans cette piste est intitulée *Cumulative Citation Recommendation (CCR)* ou encore *Vital filtering*. La tâche est motivée par le besoin de mettre à jour les bases de connaissances du genre Wikipedia. En effet, Frank *et al.* (2012) ont analysé des articles Wikipedia relatifs à des entités non populaires. Comme nous l'avons souligné, ils ont remarqué que le temps de latence médian entre la citation d'un document dans l'article Wikipedia et la date de publication de ce document est de 365 jours. Ce temps de latence est très grand rendant plusieurs articles Wikipedia obsolètes. La tâche vise à concevoir des systèmes qui analysent un flux de documents Web (News, Blogs, Forums, ...) et détectent ceux qui sont pertinents par rapport aux entités de la base de connaissances. Les documents détectés sont recommandés en temps réel (dès leur publication) aux éditeurs de la base de connaissances qui les analysent manuellement et procèdent à la mise à jour des articles avec les nouvelles informations sur l'entité.

Depuis 2013, la tâche *CCR* distingue les deux types de documents pertinents par rapport à une entité : les documents vitaux et les documents utiles.

La méthodologie d'évaluation officielle consiste à annoter chaque document du flux sur une échelle de quatre points (*Vital*, *Useful*, *Neutral* et *Garbage*). Deux types d'évaluation peuvent être effectués. Le première considère uniquement les documents vitaux comme l'ensemble de documents positifs de la vérité de terrain, alors que dans le deuxième cas d'évaluation, tous les documents pertinents (l'union des documents vitaux et utiles) constituent les documents positifs de la vérité de terrain.

Un système de filtrage *CCR* doit prédire pour chaque document sa classe (*Vital, Useful, Neutal ou Garbage*) avec un score de confiance dans l'intervalle $]1, 1000]$ tel que 1000 correspond au score de confiance le plus

fort, c'est-à-dire que le système estime que le document a une très forte chance d'appartenir à la classe prédite.

Les mesures d'évaluation adoptées dans la tâche *CCR* sont les mesures basiques telles que la *précision*, le *rappel* et la F_{mesure} (Voir la sous-section 2.2.4.2) et l'utilité normalisée (décrite dans la sous-section 2.3.2.1). La mesure d'évaluation officielle est la valeur maximale de F_{mesure} obtenue parmi tous les points de *cutoff* de la confiance entre 1 et 1000 (voir la sous-section ***hF1***).

En observant les mesures d'évaluation utilisées, nous pouvons dégager deux principales critiques :

- Malgré l'aspect temporel de la tâche *CCR*, l'évaluation officielle n'est pas sensible aux caractéristiques dépendant du temps du système (Kenter *et al.*, 2015; Dietz *et al.*, 2013). D'une part, la performance d'un système peut se dégrader extrêmement en passant de la période d'apprentissage à la période de test. D'autre part, un système peut bien fonctionner en moyenne, mais sa performance peut se dégrader lors des événements importants.
- La mesure officielle *hF1* considère le point de *cutoff* pour lequel le système est le plus performant. Cette approche rend la comparaison détaillée entre les systèmes très difficile car la stratégie d'attribution des scores de confiance peut varier d'un système à un autre ainsi que le point de *cutoff*.

Quelques auteurs se sont intéressés à résoudre ces biais (Dietz *et al.*, 2013; Kenter *et al.*, 2015). Ils ont proposé de nouveaux paradigmes d'évaluation capables d'identifier la dégradation de la performance d'un système. Ces paradigmes permettent aussi de comparer les systèmes sans la nécessité de fixer un point *cutoff* de confiance. Ils proposent de mesurer la performance des systèmes dans des périodes de temps. La différence entre ces deux propositions réside dans la manière de moyenner les performances d'un système dans les périodes de temps.

Les approches décrites dans cette section visent à filtrer les documents centrés sur une entité. Malgré leur utilité, elles renvoient des documents entiers. Dans le contexte de mise à jour des bases de connaissances, les éditeurs sont obligés de parcourir tout le contenu des documents pertinents recommandés pour chercher les nouvelles informations vitales. En outre, ces approches ne traitent pas le problème de redondance entre les documents, c'est à dire qu'elles renvoient tous les documents pertinents même s'ils contiennent des informations équivalentes.

3.4.3 Résumé temporel de documents centré sur une entité

3.4.3.1 Principe

Un système de filtrage de documents centré sur une entité analyse un flux continu et sélectionne des documents entiers potentiellement pertinents par rapport à l'entité considérée. Si nous considérons le cas où un événement important se produit ou le cas où une information nouvelle sur une entité apparaît, plusieurs documents pertinents vont apparaître pour décrire les différentes informations relatives à ces nouvelles. Tous ces documents devraient être renvoyés par un système traditionnel de filtrage de documents. L'utilisateur se trouvera alors submergé par ce grand nombre de documents renvoyés. Cette manière de filtrer a ainsi plusieurs limites, parmi lesquelles nous pouvons principalement mentionner :

- **Dispersion des informations** : Les systèmes de filtrage de documents traditionnels renvoient au cours du temps les documents potentiellement pertinents par rapport au profil défini par l'utilisateur. Les informations pertinentes concernant le profil ne sont pas toujours contenues dans un seul document. L'utilisateur doit alors analyser tous les documents reçus afin de collecter toutes les informations qui l'intéressent pour satisfaire son besoin.

- **Manque de *focus*** : les système de filtrage de documents traditionnels renvoient la totalité des documents. L'utilisateur doit scruter chaque document et chercher la partie utile qui l'intéresse.

- **Redondance d'information** : Les systèmes de filtrage de documents traditionnels renvoient des documents potentiellement pertinents sans tenir compte de la redondance qui peut exister entre eux. Si deux documents sont très pertinents par rapport au sujet d'intérêt et reportent exactement les mêmes informations, ils seront sélectionnés tous les deux. Le gain d'information apporté par un des deux documents est minime voire nul.

Le résumé temporel de documents s'intéresse à résoudre ces limites en renvoyant à l'utilisateur un résumé court, pertinent et couvrant les différentes informations sur le sujet d'intérêt tout en évitant au maximum la redondance. Ce paradigme peut avoir plusieurs applications, nous en citons ci-dessous deux exemples qui illustrent nos propos :

- ***Résumé en temps réel des microblogs*** : Cette tâche consiste à analyser un flux de microblogs (ex. *tweets*) et produire un résumé composé des microblogs pertinents permettant de couvrir les

principales informations sur un sujet donné (Sharifi *et al.*, 2010; Long *et al.*, 2011; Olariu, 2013; Mackie *et al.*, 2014).

- ***Résumé temporel d'événements*** : Le but de cette tâche est de permettre aux utilisateurs de surveiller des événements largement connus comme les catastrophes naturelles. Dans ce genre de situations de crise, les utilisateurs ont un besoin urgent d'informations surtout s'ils sont directement touchés par l'événement. Cette tâche consiste à extraire à partir d'un flux de documents, les phrases pertinentes composant ainsi un résumé temporel sur un événement donné (Aslam *et al.*, 2014).

3.4.3.2 Travaux connexes

Le résumé temporel de documents est un paradigme qui peut être lié à plusieurs domaines de recherche tels que le résumé automatique des documents et la détection de la nouveauté. Nous décrivons dans cette section, quelques travaux s'intéressant à ces domaines ainsi que les différentes approches proposées pour la génération de résumés temporels.

3.4.3.2.1 Résumé automatique de documents

Les systèmes de résumé automatique multi-documents ont pour but de produire un résumé sur un sujet donné à partir d'un ou de plusieurs documents donnés en entrée. Nenkova *et al.* (2011) ont distingué différents types de résumés automatiques parmi lesquels :

- le résumé d'un document **unique** (Hovy et Lin, 1998a),
- le résumé **multi-documents** (Maña López *et al.*, 2004; McKeown *et al.*, 2005)
- le résumé **extractif** (Erkan et Radev, 2004; Filatova et Hatzivassiloglou, 2004; Hovy et Lin, 1998b) généré en concaténant les phrases importantes extraites telles qu'elles sont écrites dans le(s) document(s),
- le résumé par **abstraction** (Ganesan *et al.*, 2010; Carenini et Cheung, 2008) composé de nouvelles phrases qui sont générées à partir des documents par des techniques de compression ou de fusion de phrases,
- le résumé de **mise à jour** (*Update Summarization*). Supposons qu'un utilisateur avait déjà lu un résumé initial sur un sujet d'intérêt, le résumé de mise à jour reporte uniquement les nouvelles informations au delà de ce que l'utilisateur connaît déjà en lisant le premier résumé (Dang et Owczarzak, 2008; Wang et Li, 2010).

Aujourd'hui, plusieurs méthodes de résumé automatique ont été proposées, les plus largement utilisées sont basées sur la technique de regroupement (*clustering*) (Radev *et al.*, 2004; Lin et Hovy, 2002; Wang

et al., 2008) ou sur la théorie des graphes (Erkan et Radev, 2004; Wan et Yang, 2008).

Quelles que soient les techniques utilisées, la quasi-totalité des travaux de résumé automatique multi-documents fonctionnent de manière rétrospective en traitant les documents comme un seul lot pertinent. Ces méthodes ne sont pas appropriées lorsqu'il s'agit de résumer des documents qui arrivent dans un flux continu. Une alternative possible consiste à résumer itérativement les nouveaux documents et fusionner le nouveau résumé avec les résumés passés. Cependant, ce type de solution soulève la question de redondance, qui est aussi difficile à traiter. Plusieurs travaux se sont intéressés à cette problématique de la redondance. Dans la sous-section suivante, nous présentons les principales méthodes proposées pour la détection de la nouveauté (non redondance) dans les textes.

3.4.3.2.2 Détection de la nouveauté

La nouveauté dans les textes a été étudiée dans plusieurs travaux de recherche d'information. Les premières recherches relatives à la détection de la nouveauté se sont intéressées à la détection et au suivi d'événements (*Topic Detection and Tracking* ou TDT en anglais) (Allan *et al.*, 1998a). Un système *TDT* surveille un flux de documents ordonnés chronologiquement, habituellement des actualités (*news stories*). Spécifiquement, la tâche de la détection de la première histoire intitulée en anglais *First Story Detection* (FSD) a pour but de détecter la première actualité qui traite d'un événement auparavant inconnu. Un événement est défini comme "quelque chose" qui arrive à un moment et lieu précis. Les approches proposées dans le cadre de la tâche FSD se basent sur un algorithme regroupement en ligne (online clustering algorithm) (Allan *et al.*, 1998b; Franz *et al.*, 2001; Spitters et Kraaij, 2001). Les nouvelles actualités sont comparées à des groupes d'actualités (*clusters*) sur des événements déjà connus. Si le nouveau document d'actualité correspond à un groupe existant, alors il décrit un événement connu, sinon il décrit un nouvel événement.

La campagne d'évaluation TREC s'est aussi intéressée au problème de détection de la nouveauté en menant la tâche *TREC Novelty* pendant les années 2002, 2003 et 2004 (Harman, 2002; Soboroff et Harman, 2003; Soboroff, 2004). Le but est d'éviter de restituer les informations redondantes par rapport à des informations déjà consultées par l'utilisateur. L'unité d'information adoptée dans la tâche est la phrase. La tâche de base est définie comme suit : étant donné un ensemble de documents pertinents, ordonnés chronologiquement et segmentés en phrases, le système doit renvoyer les phrases qui sont à la fois pertinentes et nouvelles (l'information reportée n'a pas été déjà consultée par l'utilisateur). Principalement, deux défis doivent

être surmontés : le premier est comment identifier les phrases pertinentes ? et le deuxième comment identifier à partir des phrases pertinentes celles qui contiennent des informations nouvelles ?

Plusieurs méthodes statistiques et linguistiques ont été proposées (Soboroff et Harman, 2005). Divers méthodes ont été adoptées pour détecter les phrases pertinentes, comme l'application des modèles de recherche d'information traditionnels tels que le modèle vectoriel avec la pondération *TF.IDF* (Allan *et al.*, 2003) et le modèle BM25 (Zhang *et al.*, 2004). Certaines méthodes ont utilisé la technique d'expansion du sujet ou des documents (Zhang *et al.*, 2002a). Kazawa *et al.* (2002) ont utilisé l'algorithme SVM (Machine à Vecteurs Support) pour extraire les phrases pertinentes.

Concernant la détection de la nouveauté, la méthode proposée dans (Larkey *et al.*, 2002) calcule une mesure de nouveauté en se basant sur le comptage simple des mots nouveaux n'apparaissant pas dans les documents (ou phrases) pertinents déjà restitués à l'utilisateur. Dans (Zhang *et al.*, 2002b), d'autres mesures ont été proposées, en se basant sur le calcul de la distance de cosinus entre le nouveau document en cours (d_t) et les documents déjà consultés par l'utilisateur (DT), ou sur le calcul de la divergence entre le modèle de langue de d_t et le modèle de langue de DT. Kazawa *et al.* (2002) sélectionnent les phrases nouvelles en se basant sur la mesure de la pertinence marginale maximum (MMR).

D'une façon générale, la plupart des approches proposées se sont focalisées sur la définition d'une mesure de similarité (ou distance) qui est utilisée pour comparer chaque nouveau document (article d'actualité, post, tweet, ...) entrant dans le flux à un ensemble des documents déjà vus. Si la similarité est inférieure à un seuil, le document est considéré comme nouveau. Les fonctions de similarité utilisées dans la littérature varient selon leur efficacité et leur complexité. Karkali *et al.* (2014) ont proposé une nouvelle méthode pour la détection de la nouveauté qui exploite uniquement les statistiques de flux conservées en mémoire et se base sur la définition d'une nouvelle variante du facteur *IDF* considérant l'aspect temporel. L'avantage de cette méthode est qu'elle est plus rapide et peut passer à l'échelle plus facilement.

3.4.3.3 Approches de résumés temporels centrés sur des entités

Le résumé temporel centré sur une entité est un nouveau paradigme de recherche. Son but est d'extraire à partir d'un flux de documents, les phrases pertinentes et nouvelles permettant ainsi de produire une synthèse sur une entité donnée. Dans ce cadre, la tâche *Temporal Summarization* (TS) a été lancée depuis 2013 (Aslam *et al.*, 2013, 2014). Elle s'intéresse particulièrement aux entités de type événement telles que les catastrophes naturelles, les accidents, etc.

Les approches de résumés temporels proposées dans le cadre de la tâche TS distinguent deux étapes.

La première étape concerne le filtrage des documents pertinents pour l'événement (décrit par un ensemble de mots clés formant la requête). Les approches proposées ont exploité des modèles de recherche d'information classiques tels que le modèle de langue (Baruah *et al.*, 2013), le modèle de BM25 (Yang *et al.*, 2013; Zhao *et al.*, 2014), le TF.IDF (Xu *et al.*, 2013b), le modèle booléen appliqué sur les titres des documents d'actualité (Liu *et al.*, 2013a) ou sur tout le contenu des documents (Xi *et al.*, 2013; Xu *et al.*, 2013b; Kedzie *et al.*, 2014).

La deuxième étape concerne la sélection des phrases pertinentes et nouvelles (non redondantes).

Certaines approches sélectionnent les phrases pertinentes en calculant un score basé sur les poids de mots importants relatifs à l'événement qui se trouvent dans la phrase (Liu *et al.*, 2013a; Zhang *et al.*, 2013; Chen *et al.*, 2014). Ces mots importants peuvent être récupérés de différentes façons. Liu *et al.* (2013a) s'appuient sur des documents d'apprentissage sélectionnés manuellement pour apprendre les mots importants. Zhang *et al.* (2013) ainsi que Chen *et al.* (2014) utilisent l'algorithme *Latent Dirichlet Allocation (LDA)* pour sélectionner et pondérer les mots les plus représentatifs.

Afin de conserver uniquement les phrases nouvelles, chaque phrase pertinente est comparée à toutes les phrases pertinentes déjà sélectionnées. Xi *et al.* (2013) ont utilisé la similarité de cosinus standard, alors que Liu *et al.* (2013a) ont utilisé une mesure de similarité modifiée qui renforce le poids des valeurs numériques qui sont supposées avoir une grande importance vues qu'elles peuvent évoluer au cours du temps, comme le nombre de victimes lors d'un séisme.

D'autres méthodes se sont basées sur des techniques d'apprentissage et de regroupement (*clustering*) pour sélectionner les phrases pertinentes et nouvelles. Xu *et al.* (2013a) utilisent un classifieur afin de détecter les phrases à la fois pertinentes et nouvelles. La méthode exploite différents critères mesurant la pertinence du document et de la phrase ainsi que d'autres critères qui exploitent les poids de certains termes particuliers tels que les noms d'entités, les prédicats (verbes) et les valeurs numériques. McCreadie *et al.* (2014b) ont proposé un modèle de régression exploitant plus de 300 critères décrivant les aspects prévalence, nouveauté et qualité des phrases.

Kedzie *et al.* (2014) ont utilisé dans un premier temps un classifieur pour déterminer l'importance des phrases en se basant sur des *critères de langue* mesurant la qualité et la spécificité du modèle de langue de la phrase, des *critères géographique* mesurant la distance entre les lieux identifiés dans la phrase et le lieu de l'événement (requête), des *critères*

temporels mesurant l'effet de rafales et d'autres critères de base tels que la longueur de la phrase et le nombre de mots de la requête qui apparaissent dans la phrase. Les phrases prédites comme pertinentes par le classifieur sont ensuite regroupées en utilisant un algorithme de propagation d'affinité (Dueck et Frey, 2007). Les phrases centrales (centres des clusters) seront ajoutées dans le résumé temporel.

L'évaluation de ces approches est basée sur le gain d'information acquis par l'utilisateur par chaque phrase reçue. Nous avons détaillé les mesures utilisées dans l'évaluation dans la sous-section 2.3.2.3.

3.5 Conclusion

Les entités du monde réel peuvent évoluer de manière régulière engendrant l'apparition de plusieurs documents reportant de nouvelles informations sur ces entités, et rendant obsolètes les anciens documents ou les articles descriptifs de ces entités dans les bases de connaissances.

Dans ce chapitre, nous avons présenté les approches de l'état de l'art s'intéressant au filtrage et à l'agrégation d'informations autour d'entités. Les informations filtrées peuvent être présentées sous forme de **documents** ou d'un **résumé temporel**.

Les approches de filtrage de documents centrés sur une entité renvoient des documents entiers ce qui rend difficile la tâche de l'utilisateur qui doit chercher les informations qui l'intéressent. De plus, ces approches ne traitent pas le problème de redondance. Les approches de résumés temporels se sont intéressées à ce problème en visant à extraire uniquement les phrases importantes et nouvelles.

Dans cette thèse, nous nous intéressons particulièrement à :

- la **détection des documents vitaux**. Nous proposons deux méthodes de détection de documents vitaux. La première est totalement *supervisée* et dépend de la présence de documents d'apprentissage pour chaque entité, mais l'avantage est qu'elle exploite uniquement un seul critère (facteur) de vitalité. La deuxième méthode proposée est originale et *non supervisée*. Elle exploite un nouveau facteur temporel de vitalité basé sur la reconnaissance des dates dans le texte du document.

- l'**agrégation d'informations vitales issues des documents détectés**. Nous proposons une méthode qui exploite les annotations associées aux entités similaires dans une base de connaissances afin d'en apprendre les mots importants que nous utilisons pour sélectionner les phrases vitales. Pour détecter la redondance des phrases,

nous proposons une fonction de nouveauté qui combine la divergence textuelle avec la détection des nouvelles entités liées.

Deuxième partie : Filtrage et agrégation d'informations vitales autour d'une entité

Introduction

Cette partie de ce mémoire, intitulée ***filtrage et agrégation d'informations vitales autour d'une entité***, présente nos contributions. Nos travaux visent à filtrer et agréger les d'informations vitales et nouvelles relatives à des entités. Nous souhaitons construire un résume temporel pour chaque entité d'intérêt. Réaliser cet objectif nécessite de traiter deux étapes, comme illustrées dans la figure 3.6. D'abord, nous nous intéressons au filtrage de documents vitaux à partir d'un flux. Ensuite, nous sélectionnons les phrases vitales et nouvelles (non redondantes) à partir des documents vitaux filtrés.

Concernant la première étape, nous proposons deux méthodes pour filtrer les documents vitaux décrites dans le chapitre 4.

- La première méthode est *supervisée*, c'est à dire qu'elle nécessite des documents d'apprentissage pour chaque entité. Elle se base sur les modèles de langue (Ponte et Croft, 1998; Lavrenko et Croft, 2001) qui ont montré de bonnes performances pour mesurer la notion de pertinence dans la recherche d'information. Nous voulons évaluer la capacité de ces modèles à estimer la vitalité (Abbes *et al.*, 2014b,a).
- La seconde méthode est *non supervisée*. Nous proposons d'estimer la vitalité d'un document en exploitant sa fraîcheur que nous estimons à partir des expressions temporelles reportées dans son texte (Abbes *et al.*, 2015c,d).

Afin d'évaluer nos deux approches de détection de documents vitaux, nous nous sommes basés sur les corpus de documents fournis par la campagne d'évaluation TREC (Text Retrieval Conference) dans la tâche *Knowledge Base Acceleration* en 2013 et 2014.

Concernant la deuxième étape, nous proposons deux solutions pour résoudre les deux problématiques relatives à la *sélection des phrases vitales* et la *détection de la nouveauté (non redondance)* (Abbes *et al.*, 2015b,a).

- D'abord, nous proposons de sélectionner les phrases vitales en nous basant sur la présence de mots importants (déclencheurs) récupérés automatiquement en exploitant le vocabulaire propre aux entités

similaires ayant le même type que l'entité considérée dans une base de connaissance (dans notre cas, *DBpedia*).

- Ensuite, nous proposons une méthode qui détecte la nouveauté (non redondance) en combinant la divergence textuelle avec l'identification des entités liées dans les phrases.

Nous combinons ces deux solutions afin de générer un résumé temporel sur l'entité.

Pour évaluer les méthodes proposées dans cette étape (génération d'un résumé temporel), nous nous sommes basés sur les corpus de documents fournis par la campagne d'évaluation TREC dans la tâche *Temporal Summarization* en 2013 et 2014.

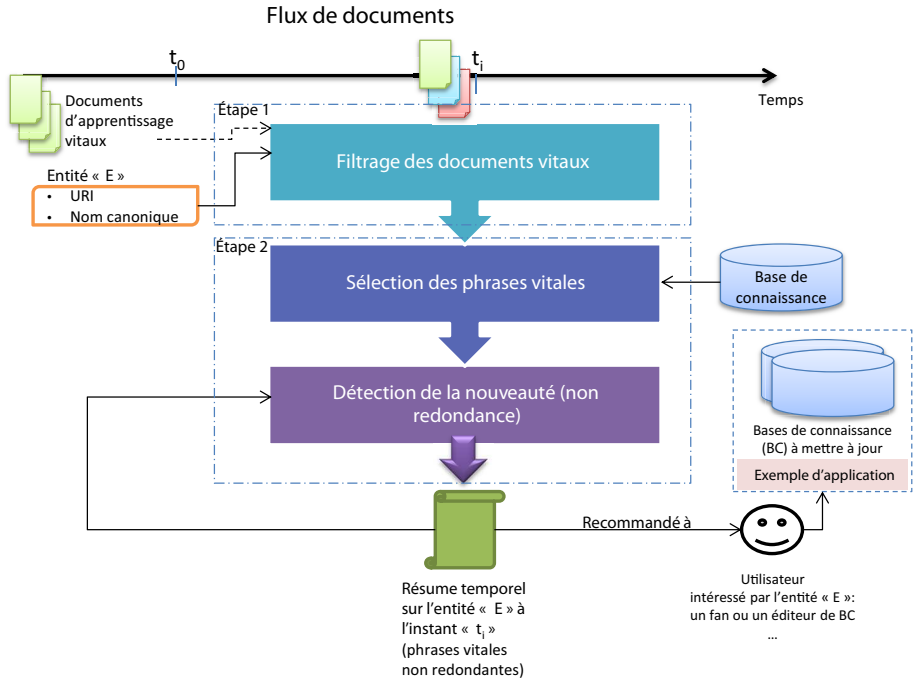

FIGURE 3.6 – Architecture de note système de filtrage et d'agrégation d'informations vitales relatives à une entité

Chapitre 4

Filtrage de documents vitaux autour d'une entité

4.1 Introduction

Nous décrivons dans ce chapitre nos premières propositions relatives au filtrage de documents centrés sur une entité. Nous rappelons que le but est de renvoyer à partir d'un flux continu de documents ceux qui sont pertinents par rapport à une entité donnée. Dans ce chapitre, nous prenons la tâche de *TREC Cumulative Citation Recommendation (CCR)* comme un exemple d'application. Dans ce cadre, deux classes de pertinences peuvent être distinguées : les documents utiles (*useful documents*) et les documents vitaux (*vital documents*). Les documents utiles fournissent des informations pertinentes sur l'entité mais n'apportent pas de nouvelles informations. Ils peuvent être intéressants si nous désirons construire une page descriptive de l'entité en partant de rien. Cependant, ils seront moins intéressants lorsqu'il s'agit d'enrichir une page descriptive existante. Les documents vitaux reportent des informations pertinentes et nouvelles permettant de maintenir de façon continue la mise à jour des pages descriptives des entités dans les bases de connaissances. C'est pour cette raison que nous nous intéressons particulièrement dans ce travail au filtrage des documents ***vitaux***.

Nous illustrons dans la figure 4.1, l'architecture générale de note système de filtrage de documents vitaux. Le système est alimenté par une entité d'intérêt décrite par son nom canonique et l'URI de sa page descriptive (par exemple sa page Wikipedia). Le processus de filtrage implique deux phases : une "*phase de filtrage*" dans laquelle nous devons décider si un document arrivant doit être pris comme candidat ou non, suivie d'une "*phase d'attribution de score*" au document.

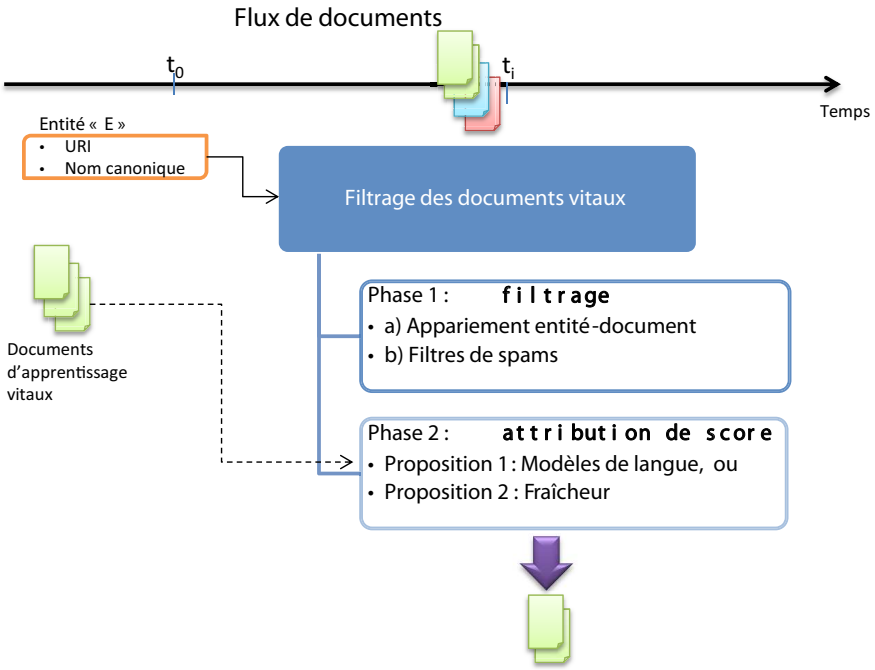

FIGURE 4.1 – Architecture de note système de filtrage de documents vitaux

Dans ce travail, pour la *phase de filtrage*, nous souhaitons éliminer plusieurs documents non pertinents du flux. Pour cela, nous considérons les documents mentionnant l'entité. Ensuite, nous appliquons des filtres de spams éliminant les documents qui ont tendance à être spams. Nous détaillons cette phase dans la section 4.4.2.1.

Concernant la *phase d'attribution de score*, nous proposons deux méthodes.

- La première méthode, que nous décrivons dans la section 4.2, est supervisée (Abbes *et al.*, 2014b). Elle exploite un seul facteur basé sur un modèle de langue.
- La deuxième méthode, décrite dans la section 4.3, est non supervisée (Abbes *et al.*, 2015c,d). Elle combine deux facteurs : un facteur de pertinence thématique basée sur le modèle de langue, et un facteur de fraîcheur exploitant les dates reconnues dans le document.

Pour valider ces deux propositions, nous menons des expérimentations dans le cadre de la tâche *Cumulative Citation Recommendation* proposée dans la campagne d'évaluation TREC en 2013 et 2014.

4.2 Proposition 1 : Modèles de langues pour la détection de la vitalité

Notre but est d'identifier à partir d'un flux de documents ceux qui sont vitaux par rapport à une entité donnée. Intuitivement, nous supposons qu'un document vital utilise un ensemble de termes qui peuvent refléter la vitalité. Nous mettons l'hypothèse que lorsque ces termes apparaissent dans un documents récent (qui vient d'être publié dans le Web), ce dernier tend à être vital.

Par exemple, entre un document annonçant la sortie de l'album `Sans attendre (2012)` de `Celine Dion` et un document annonçant la sortie son dernier album `Loved Me Back To Life (2013)`, nous pouvons trouver des termes en commun comme `dévoiler`, `publier`, `nouveau`, `celinedion.com`, etc. Notre idée est de détecter ces termes à partir des anciens documents vitaux, ceux annonçant la sortie de l'album `Sans attendre` ; nous pensons que ces termes "annonciateurs de nouveauté" nous allons les retrouver dans les nouveaux documents vitaux, ceux annonçant la sortie de l'album `Loved Me Back To Life`.

Dans la figure 4.2 qui concerne l'entité `Bertrand Monthubert`, nous pouvons remarquer certains termes qui se répètent entre les document utiles (documents `E` et `F`, qui peuvent être vus comme des anciens des documents vitaux) et les nouveaux nouveaux vitaux (documents `B`, `C` et `D`). Parmi ces termes, nous trouvons : `rapport`, `national`, `enseignement`, `supérieur`, `démissionner`, `www.letudiant.fr`.

Le modèle de langue (Ponte et Croft, 1998) que nous avons décrit dans la section 2.2.2.3 permet de modéliser un texte (un ou plusieurs documents) en capturant la distribution des mots. Lavrenko et Croft (2001) ont proposé un modèle de langue de pertinence pour estimer les probabilités des mots dans une classe pertinente. Dans notre approche, nous considérons que la vitalité représente une classe de pertinence qui peut être modélisée, par analogie au modèle de pertinence, par un modèle de langue que nous nommons *modèle de vitalité*.

Ce modèle de vitalité vise à capturer les termes "vitaux" permettant d'identifier les documents vitaux à venir. Comme la vitalité est inconnue a priori, nous supposons que nous pouvons tirer parti d'un ensemble de documents échantillons vitaux du passé. Bien que les documents vitaux du passé puissent devenir "utiles" (non vitaux) aujourd'hui, nous pensons qu'ils peuvent aider à détecter de nouveaux documents vitaux.

Estimer un modèle de vitalité est une tâche complexe qui nous pousse à

FIGURE 4.2 – Exemples de documents vitaux et utiles étant donné la page Wikipedia de l'entité `Bertrand Monthubert`

poser plusieurs hypothèses :

1. En termes de **généralité**, nous considérons à tour de rôle les deux hypothèses suivantes :
   - *Hypothèse 1 :* La vitalité est indépendante de l'entité. Elle peut ainsi être estimée par un modèle de vitalité général qui permettra de détecter les documents vitaux de n'importe quelle entité.
   - *Hypothèse 2 :* La vitalité est dépendante de l'entité. Par conséquent, chaque entité doit avoir un modèle de vitalité spécifique qui servira

à détecter ses documents vitaux.

2. En termes d'**unité** d'estimation, nous considérons également à tour de rôle les deux hypothèses suivantes :
   - Hypothèse 3 : Tout le contenu d'un document vital est nécessaire pour estimer le modèle de vitalité.
   - Hypothèse 4 : Une seule partie de document (phrases, paragraphes) est suffisante pour estimer le modèle de vitalité.

3. En termes de **dimension**, nous considérons les deux hypothèses suivantes :
   - Hypothèse 5 : Nous considérons la vitalité comme *unidimensionnelle*. Nous pouvons alors estimer un seul modèle de vitalité à partir d'un seul document "virtuel" qui représente la concaténation de tous les documents échantillons vitaux. Ce modèle considère que ces échantillons ont la même importance. Nous détaillons l'estimation unidimensionnelle de la vitalité dans la section 4.2.1.
   - Hypothèse 6 : Nous considérons la vitalité comme *multidimensionnelle*. Nous considérons que les documents échantillons vitaux n'ont pas tous la même importance. Pour cela nous estimons un modèle par document. Ensuite, nous estimons un seul modèle de vitalité en agrégeant et pondérant les modèles de vitalité des documents. Nous détaillons l'estimation multidimensionnelle de la vitalité dans la section 4.2.2.

Dans la section 4.4.5, nous expérimentons les hypothèses posées afin de répondre aux questions suivantes :

- Est-il possible d'estimer un modèle de vitalité général permettant de détecter des documents vitaux de n'importe quelle entité ?
- La pondération des documents échantillons vitaux permet-elle d'améliorer l'estimation du modèle de vitalité ?
- Est-il nécessaire de considérer tout le contenu d'un document afin d'estimer un modèle de vitalité performant ?

4.2.1 Estimation d'un modèle de vitalité unidimensionnel

Selon l'hypothèse 5 que nous avons posée précédemment, nous considérons la vitalité comme *unidimensionnelle*. Nous considérons que tous les documents d'apprentissage ont la même importance. Ils contribuent à estimer la vitalité de manière uniforme. Par conséquent, nous voyons les documents échantillons vitaux comme un seul document "virtuel" construit par concaténation.

Formellement, nous supposons disposer d'un ensemble de documents d'apprentissage vitaux. Nous notons cet ensemble vital V =

$\{dv_1, dv_2, \ldots, dv_m\}$, tel que dv_i représente tout le contenu ou une partie d'un document échantillon vital.

Dans la suite, nous parlerons de document échantillon vital dv_i, dv_i pour faire référence à tout ou une partie d'un document échantillon vital.

Soit DV la représentation d'un document virtuel concaténant tous les documents échantillons vitaux appartenant à l'ensemble V. Nous estimons la probabilité de générer un terme t à partir d'un modèle de vitalité unidimensionnel θ_{V_u} en utilisant le lissage de Dirichlet (MacKay et Peto, 1994) :

$$P(t|\theta_{V_u}) = \frac{tf(t, DV) + \mu\ P(t|C)}{|DV| + \mu} \tag{4.1}$$

où

- C est une collection de référence comportant des documents antérieurs dans le flux.
- $tf(t, DV)$ représente la fréquence d'apparition du terme t dans le document DV
- μ représente une valeur de lissage réelle $\in [0, +\infty[$
- $P(t|C) = \frac{tf(t,C)}{\sum_{t' \in T} tf(t',C)}$, avec T représente tous les termes du vocabulaire.

L'estimation unidimensionnelle peut être effectuée de manière **spécifique** à une entité spécifique E en considérant DV comme la concaténation de tous les documents échantillons vitaux relatifs à cette entité (hypothèse 2). Comme elle peut être **générale** (indépendante de l'entité) en considérant DV comme la concaténation de tous les documents échantillons vitaux disponibles pour des entités différentes (hypothèse 1).

4.2.2 Estimation d'un modèle de vitalité multidimensionnel

Selon l'hypothèse 6, nous considérons la vitalité comme *multidimentionnelle*. Nous pensons ici que les documents vitaux d'apprentissage n'ont pas tous la même importance. Pour cela nous estimons un modèle de vitalité par document. Ensuite, nous estimons un seul modèle de vitalité qui agrège et pondère les modèles de vitalité.

Soit E_i une entité ayant m_i documents échantillons vitaux dv_{ij}. Par analogie au modèle de pertinence proposé par Lavrenko et Croft (2001), la probabilité de générer un terme t à partir d'un modèle de vitalité multidimensionnel $\theta_{V_{m_i}}$ spécifique à l'entité E_i est estimée comme suit :

$$
\begin{aligned}
P(t|\theta_{V_{m_i}}) &= \sum_{j=1}^{m_i} P(t|\theta_{dv_{ij}}) \frac{P(\theta_{dv_{ij}}|E_i)}{\sum_{k=0}^{m_i} P(\theta_{dv_{ik}}|E_i)} \\
&\propto \sum_{j=1}^{m_i} P(t|\theta_{dv_{ij}}) \frac{P(E_i|\theta_{dv_{ij}})}{\sum_{k=0}^{m_i} P(E_i|\theta_{dv_{ik}})}
\end{aligned}
\tag{4.2}
$$

où

- m_i représente le nombre de documents échantillons vitaux pour l'entité E_i.

- $P(E_i|\theta_{dv_{ij}})$ est estimée par l'équation 4.3.

$$
P(E_i|\theta_{dv_{ij}}) = \prod_{q \in nom\_canonique(E_i)} P(q|\theta_{dv_{ij}}) \tag{4.3}
$$

avec

- $nom\_canonique(E_i)$ est le nom canonique de l'entité E_i composé de termes q.
- $P(q|\theta_{dv_{ij}})$ est l'estimation d'un modèle de vitalité unidimensionnel pour l'entité E_i à partir d'un seul document vital dv_{ij}. Cette probabilité est calculée en utilisant le lissage de Dirichlet de manière similaire à l'équation 4.1.

Nous pouvons déduire un modèle multidimensionnel général $\theta_{V_{m_G}}$ indépendant de l'entité. Nous supposons avoir n entités pour lesquelles nous disposons d'un ensemble de documents d'apprentissage vitaux. Pour chaque entité E_i, nous utilisons l'équation 4.2 pour estimer le modèle multidimensionnel spécifique θV_{m_i}. Le modèle multidimensionnel général est donné par l'équation suivante :

$$
P(t|\theta_{V_{m_G}}) = \frac{1}{n} \sum_{i=1}^{n} P(t|\theta_{V_{m_i}}) \tag{4.4}
$$

4.2.3 Mesure de la vitalité d'un document basée sur un modèle de vitalité

Le modèle de vitalité estimé à partir d'un ensemble de documents d'apprentissage servira à détecter de nouveaux documents vitaux. L'évaluation de la vitalité d'un nouveau document d par rapport à une entité E est traduite par le calcul d'un score de vitalité mesurant une certaine similarité entre le modèle du document θ_d et le modèle de vitalité appris θ_{V_x} (unidimensionnel/multidimensionnel, général/spécifique).

Dans ce travail, nous estimons cette similarité par le calcul de la vraisemblance des termes du modèle vital θ_{V_x} dans le document d comme

donnée par l'équation suivante :

$$Score_{Vitalité}(d, E) = \prod_{t \in top(\theta V_x, k)} P(t|\theta_d)^{P(t|\theta_{V_x})} \tag{4.5}$$

Avec,
- $P(t|\theta_d)$ est calculée avec le lissage de Dirichlet.
- Nous considérons uniquement le top k des termes représentatifs dans le modèle θ_{V_x}
- $P(t|\theta_{V_x})$ est utilisée pour pondérer les termes du modèle vital θ_d.

Dans les expérimentations, afin d'étudier l'impact de l'unité d'information choisie dans la détection de la vitalité (hypothèses 3 et 4), nous estimons le modèle de document θ_d et le modèle de vitalité θ_{V_x} en considérant trois unités d'informations (phrases mentionnant l'entité E, les paragraphes mentionnant l'entité E et tout le contenu) des documents d'apprentissage ou d'évaluation.

4.3 Proposition 2 : Exploitation des expressions temporelles pour la détection de la vitalité

Dans cette section, nous proposons une deuxième approche totalement non supervisée. Elle ne nécessite aucun document d'apprentissage. Notre idée est de considérer qu'un document vital doit être récent. Nous supposons avoir une date référence dénotée par t_0. Nous considérons qu'un document est récent s'il est publié après la date de référence t_0. Par exemple, dans le cadre de la tâche CCR, la date de référence t_0 correspond à la date de la dernière mise à jour de la page descriptive de l'entité à mettre à jour.

Dans cette approche, nous pensons que la fraîcheur représente un critère essentiel qui joue un rôle important pour estimer la vitalité d'un document. Ce critère peut être déterminé en vérifiant la date de publication du document et les expressions temporelles utilisées dans son texte. Nous posons l'hypothèse suivante : plus les dates inférées par les expressions temporelles sont proches de la date de publication, meilleure est la tendance du document à être frais et par conséquent à être vital pour l'entité.

Dans la figure 4.2, les documents `'D'`, `'E'` et `'F'` (publiés en `Septembre, 2015`) sont pertinents par rapport à l'entité `Bertrand Monthubert (mathématicien et homme politique)`. Cependant, nous pouvons remarquer que le document 'E' est plus frais que le document `'D'` parce qu'il mentionne une expression temporelle référant à une date (`8 Septembre 2015`) proche à la date de publication du document, alors que le document `'D'` contient des dates anciennes (`2008`, `2012`) par rapport à sa

date de publication. La même constatation peut être faite pour le document `'F'`.

Formellement, soient une entité E et un nouveau document d ayant une date de publication $Date_p(d)$. Supposons que $Date_t^*$ est la date la plus proche de $Date_p(d)$ reconnue dans le document d (ou dans la partie du document mentionnant l'entité E). Nous supposons que plus de délai $\mathbf{\Delta(d, E) = |Date_p(d) - Date_t^*|}$ est court, plus grande est la probabilité de vitalité du document d. Nous traduisons cette intuition par un score de fraîcheur calculé selon une fonction exponentielle négative.

Lorsque le document d mentionne une expression temporelle référant à sa date du publication, le délai $\Delta(d, E)$ sera égal à 0 permettant d'atteindre la valeur maximale du score de fraîcheur (égal à 1). Lorsque le délai $\Delta(d, E)$ augmente, le score de fraîcheur tend à diminuer jusqu'à ce qu'il atteigne la valeur 0.

Le score de fraîcheur est donné par l'équation suivante :

$$Score_{Fraîcheur}(d, E) = e^{-\frac{\Delta(d,E)^2}{\sigma^2}} \tag{4.6}$$

$$\Delta(d, E) = \min_{x \in X(d,E)} (|Date_p(d) - Date_t(x, d)|) \tag{4.7}$$

avec

- $Date_p(d)$ est la date de publication du document d.
- $X(d, E)$ est l'ensemble des expressions temporelles détectées à partir des parties du document d (phrases, paragraphes, tout le contenu) qui mentionnent l'entité E.
- $Date_t(x, d)$ est la date inférée par l'expression temporelle x.
- Nous notons que lorsque la partie considérée du document d ne contient aucune expression temporelle, $Score_{Fraîcheur}(d, E)$ sera égal à 0.
- Nous utilisons le nombre de jours comme unité pour calculer le délai entre $Date_p(d)$ et $Date_t(x, d)$.

Dans cette approche, nous pensons que la fraîcheur est un critère important pour déterminer la vitalité d'un document. Cependant, elle doit être combinée avec la pertinence thématique du document. En effet, un document peut être frais sans mentionner aucune information sur l'entité. Pour modéliser la pertinence thématique de l'entité, nous supposons disposer d'une page descriptive de l'entité (par exemple sa page Wikipedia). Nous utilisons cette page pour estimer un modèle de pertinence θ_{P_E}. Nous calculons le score de pertinence d'un document d par rapport à l'entité E par la vraisemblance des termes du modèle de pertinence θ_{P_E} dans le document d comme donnée par l'équation suivante :

$$Score_{Pertinence}(d, E) = \prod_{t \in top_k(P_E)} P(t|\theta_d)^{P(t|\theta_{P_E})} \quad (4.8)$$

- $top_k(P_E)$ est l'ensemble de top k termes dans θ_{P_E}. Ce paramètre peut être déterminé expérimentalement.
- $P(t|\theta_d)$ et $P(t|\theta_{P_E})$ sont estimées en utilisant le lissage de Dirichlet

Finalement, nous calculons le score de vitalité par l'équation 4.9 comme le produit des scores de pertinence et de fraîcheur. Nous ajoutons un paramètre $\epsilon \in]0, 1]$ dans le deuxième opérande afin d'éviter d'avoir un score de vitalité nul causé par l'absence d'une expression temporelle dans le document.

$$Score_{Vitalité}(d, E) = Score_{Pertinence}(d, E) * (Score_{Fraîcheur} + \epsilon) \quad (4.9)$$

4.4 Expérimentations

Nous évaluons nos approches dans le cadre de la tâche *Cumulative Citation Recommendation (CCR)* de TREC KBA 2013 et 2014. *KBA* est une piste de TREC ayant comme but d'accélérer la mise à jour des bases de connaissances. Dans cette section, nous commençons par décrire la cadre expérimental dans lequel nous avons mené nos expérimentations. Nous présentons les corpus utilisés, les topics et la manière dont les jugements de pertinence ont été effectués. Ensuite, nous présentons les expérimentations relatives aux deux approches que nous avons proposées. Pour l'approche basée sur les modèles de langue, nous étudions les performances des différentes hypothèses posées pour estimer la vitalité. Pour l'approche basée sur l'exploitation des expressions temporelles, nous montrons l'importance du facteur fraîcheur et nous étudions l'impact de la partie de document considérée pour estimer ce facteur. Enfin, nous comparons nos deux approches par rapport aux méthodes proposées dans la tâche CCR de TREC KBA 2013 et 2014.

4.4.1 Cadre expérimental

4.4.1.1 Topics évalués dans la tâche CCR 2013 et 2014

La tâche CCR propose 141 topics en 2013 et 109 topics en 2014. Ces topics correspondent à des entités de type personne (PER), organisation (ORG) et établissement (FAC). Dans la tâche CCR 2013, chaque topic est identifiée par l'*URI* de sa page descriptive correspondant à la page Wikipedia ou au profil Twitter de l'entité. En 2014, une URI "locale" (non accessible) a été fournie. Les figures 4.3 et 4.4 montrent des exemples de topics proposés dans les tâches CCR 2013 et CCR 2014 respectivement.

L'URI d'une entité est donnée par le champ *target_id*. Pour les topics de 2014, le nom canonique de l'entité est donné explicitement par le champ *canonical_name*.

```
{
"entity_type": "FAC",
"group": "fargo",
"target_id":http://en.wikipedia.org/wiki/Fargo_Air_Museum
},
```

FIGURE 4.3 – Exemple de topic proposé dans la tâche CCR 2013

```
{
    "canonical_name": "Joby Shimomura",
    "entity_type": "PER",
    "external_profile": [
    "http://www.governor.wa.gov/office/seniorstaff.aspx"
    ],
    "target_id": "https://kb.diffeo.com/Joby_Shimomura",
    "training_time_range_end": "2013-01-03-22"
},
```

FIGURE 4.4 – Exemple de topic proposé dans la tâche CCR 2014

4.4.1.2 Corpus de TREC KBA 2013 et 2014

La campagne d'évaluation TREC a élaboré plusieurs versions de corpus [1] comportant des millions de documents issus de plusieurs sources (news, Social, Forum, Weblog, etc.). Les documents sont datés et répertoriés par heure. Nous avons utilisé les versions officielles de 2013 et 2014 selon lesquelles les jugements de pertinences ont été élaborés. Des statistiques sur ces versions sont données dans le tableau 4.1.

Les corpus sont stockés dans différents répertoires (11948 pour KBA 2013 et 13663 pour KBA 2014), chacun représentant une heure distincte. Le corpus KBA 2013 est très volumineux avec une taille de 4.5 TB, contenant environ 1 milliard de documents datant du mois d'octobre 2011 jusqu'au mois de février 2013. La version d'origine de KBA2014 étend celle de 2013 pour couvrir de nouveaux documents allant jusqu'à Mai 2013. Vue sa grande taille (1.2 milliard de documents), une version filtrée a été élaborée par les organisateurs de la tâche contenant uniquement les documents

1. http ://s3.amazonaws.com/aws-publicdatasets/trec/kba/index.html

TABLE 4.1 – Statistiques sur les corpus de TREC KBA 2013 et 2014 utilisés dans l'évaluation

Corpus	**KBA 2013**	**KBA 2014**
version	v0_2_0-english	v0_3_0-kba-filtered
Intervalle de temps		
* Début	2011-10-05-00	2011-10-05-00
* Fin	2013-02-13-23	2013-05-01-02
Sources	**Nombre de documents**	
arxiv	126 952	9 075
CLASSIFIED	14 755 278	100 669
FORUM	36 559 578	480 581
linking	5 448 875	189 445
MAINSTREAM_NEWS	57 391 714	2 819 579
MEMETRACKER	7 637	265
news	134 933 117	4 772 483
REVIEW	52 412	3 011
social	394 381 405	3 218 775
WEBLOG	396 863 627	8 900 377
Total	**1 040 520 595**	**20 494 260**
Taille (compressée)	**4.5 TB**	**639 GB**

s'appariant avec les topics proposés en utilisant les noms canoniques des entités comme requêtes. Dans les expérimentation, nous utilisons la version filtrée (colonne 2 du tableau 4.1). Cette version contient environ 20.5 millions de documents avec une taille de *639 GB*.

Nous avons utilisé la bibliothèque open source *Lucene*[2] pour indexer et rechercher des documents. Chaque répertoire (heure) est indexé dans un index à part pour faciliter la simulation du corpus en tant qu'un flux de documents.

4.4.1.3 Annotation des documents et périodes d'apprentissage et d'évaluation

En **2013**, les juges de la tâche ont annoté environ *50 000* documents qui mentionnent les noms d'entités étudiées. Les annotations des documents datant d'avant Février 2012 sont fournies en tant que données d'apprentissage. Les documents datant de Mars 2012 jusqu'à Février 2013 sont utilisés pour l'évaluation. Un document est annoté par plusieurs juges qui assignent la classe correspondante à l'entité cible (*vital, useful, neutral, garbage*). L'heure *2012-03-01-00* sépare les documents du flux en

2. https ://lucene.apache.org/

deux périodes (apprentissage et évaluation). Cette date est la même pour tous les topics. Ceci implique que certains topics n'ont pas de documents d'entraînement dans la période d'apprentissage (avant *2012-03-01-00*). Un document est annoté comme *vital* s'il reporte une information pertinente et nouvelle, *utile* (*useful*) s'il reporte une information pertinente déjà connue, *neutre* (*neutral*) s'il mentionne une information non importante sur l'entité et *déchet* (*garbage*) s'il ne fournit aucune information sur l'entité. Nous rappelons que dans ce travail, nous nous intéressons uniquement à la détection des documents vitaux. Un document est considéré comme vital s'il a été annoté comme vital par l'unanimité des juges. Il est à noter qu'à l'issue des annotations, uniquement **122** topics ont au moins un document vital (dans la période d'apprentissage ou d'évaluation).

En **2014**, les juges de la tâche ont annoté environ *30 000* documents. Contrairement à l'année 2013, la date qui sépare la période d'apprentissage et la période d'évaluation n'est plus la même pour tous les topics. Cette date est définie pour chaque topic (champ *training_time_range_end* dans la figure 4.4) de telle sorte que les documents de la vérité terrain soient divisés en deux ensembles : 20% pour l'apprentissage et 80% pour l'évaluation. Cette stratégie augmente la chance d'avoir des documents d'apprentissage pour la plupart des topics.

Une deuxième différence par rapport à l'année 2013 consiste à considérer qu'un document est vital si au moins un juge l'a annoté par cette classe (vital). A l'issue des annotations, uniquement **71** topics ont au moins un document vital dans la période d'apprentissage.

Nous récapitulons dans le tableau 4.2 quelques statistiques sur le nombre de documents vitaux dans la tâche CCR 2013 et 2014.

TABLE 4.2 – Statistiques sur les documents vitaux dans la tâche CCR 2013 et 2014

	2013	**2014**
Nombre d'entités avec au moins 1 document vital d'apprentissage	88	71
Nombre d'entités avec au moins 1 document vital de test	108	71
Nombre d'entités évaluées (ayant au moins 1 document vital de test ou d'apprentissage)	122	71
Nombre de documents vitaux d'apprentissage	1 619	584
Nombre de documents vitaux d'évaluation	3 922	3584
Nombre de documents vitaux (total)	5 541	4 168
Moyenne de documents vitaux d'apprentissage	13	8
Moyenne de documents vitaux d'évaluation	32	50

4.4.1.4 Évaluation et métrique

Nous avons utilisé l'outil d'évaluation officiel [3] donné par les organisateurs de la tâche. La mesure d'évaluation officielle adoptée dans cette tâche est la mesure ***hF1***. Plus précisément, la mesure *hF1* consiste à la moyenne harmonique maximale entre la macro-moyenne de la précision et la macro-moyenne du rappel parmi tous les points de cutoff. Cette mesure est donnée par l'équation 2.10 dans la section 2.3.2.2.

4.4.2 Démarche d'évaluation des approches de filtrage de documents vitaux sur une entité

Généralement, un système de filtrage de documents centrés sur des entités implique deux étapes principales. La première étape qu'on appelle "phase filtrage", consiste à décider si le document doit être pris comme candidat ou rejeté. La deuxième étape, appelée "phase d'attribution de score", consiste à attribuer des scores aux documents acceptés. Les documents ayant un score dépassant un seuil donné seront fournis à l'utilisateur et les autres seront rejetés.

Dans nos méthodes, nous sélectionnons d'abord les documents s'appariant avec l'entité et ne semblant pas être des spams (phase 1 : sous-section 4.4.2.1). Ensuite, nous assignons un score de vitalité à chaque document en appliquant l'un des modèles décrits dans les sections 4.2 et 4.3 (phase 2).

4.4.2.1 Phase 1 : Filtrage des documents mentionnant l'entité

Cette étape permet de décider a priori si les documents qui arrivent sont potentiellement vitaux ou non. Ce premier filtre permet de réduire le nombre de documents pour lesquels nous estimons une vitalité. Lorsqu'un document arrive, nous procédons à deux niveaux de filtrage. Tout d'abord, nous vérifions si le document mentionne l'entité considérée. Ensuite, nous vérifions si le document est informatif (n'est pas un spam).

a) Détermination des documents mentionnant l'entité

Une entité donnée peut être référencée dans un document avec différentes formes lexicales possibles appelées variantes. Nous notons ces variantes par l'ensemble $variantes(E) = \{\text{“}var''_1, \text{“}var''_2, ..., \text{“}var''_n\}$ tel que var_i est une forme lexicale possible du nom de l'entité E. Une variante var_i peut contenir un ou plusieurs termes.

Afin d'identifier tous les documents mentionnant l'entité, nous devons connaître ses différentes variantes. Pour cela, nous utilisons différentes stratégies selon la source de l'entité :

3. https ://github.com/trec-kba/kba-scorer

- pour une entité Wikipedia, nous avons exploité la page Wikipedia en extrayant le titre et les mots en gras dans le premier paragraphe comme variantes (Cucerzan, 2007). Nous considérons le titre de la page Wikipedia comme le nom canonique de l'entité. Dans la figure 4.5a, pour l'entité `http ://en.wikipedia.org/wiki/Jeff Severson`, les variantes extraites sont `Jeff Severson` (représente aussi le nom canonique) et `Jeffrey Kent Severson`.
- Pour une entité Twitter, nous avons exploité sa page Twitter en considérant son nom affiché comme la seule variante de l'entité et son nom canonique. Comme le montre la figure 4.5b, le nom canonique de l'entité `https://twitter.com/redmondmusic` est `Eric Redmond`.
- Pour les entités de la tâche CCR 2014, nous avons considéré une seule variante telle que donnée dans la définition du topic (champ *canonical_name* dans la figure 4.4).

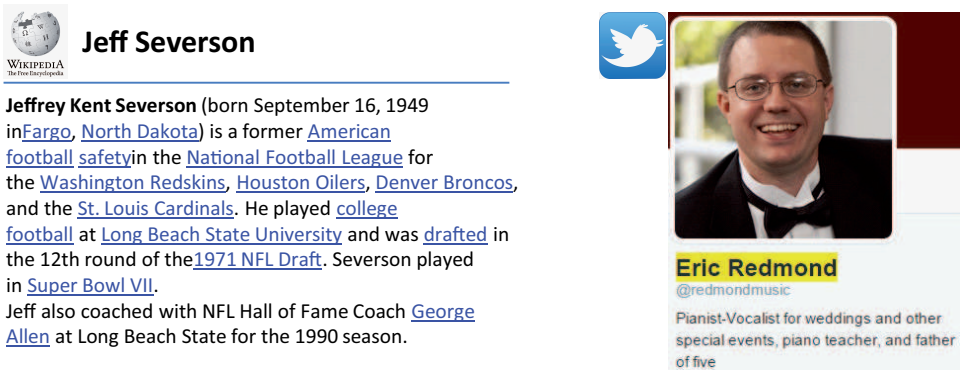

FIGURE 4.5 – Extraction des variantes d'entités à partir de Wikipedia et de Twitter

Pour déterminer si un document concerne ou pas l'entité, nous utilisons une requête booléenne pour vérifier s'il contient au moins une variante de l'entité. La requête *Lucene* soumise est la suivante :

$$requête(E) = '' var_1'' \sim 1 \; OR \; ''var_2'' \sim 1 \; OR \; ... \; OR \; ''var_n'' \sim 1$$

~ 1 permet d'effectuer une recherche à proximité relaxée acceptant au maximum l'occurrence d'un mot de plus au sein de l'expression var_i. Par exemple, en considérant la variante d'entité `"Michael Jackson"`, un document mentionnant l'expression `"Michael J. Jackson"` sera sélectionné par la requête relaxée.

b) Élimination des documents Spam
Rejeter les documents spam permet de gagner en temps de calcul ainsi que d'améliorer la performance du système. Pour cela, nous avons défini trois filtres afin de rejeter les documents s'appariant avec l'entité mais susceptibles d'être des spams.

- ***Filtre de langue*** qui élimine tous les documents reconnus comme non anglais à l'aide d'un détecteur de langue implémenté en Java [4].

- ***Filtre d'énumération*** : Nous supposons que lorsqu'une partie de document mentionne plusieurs entités de manière condensée et que lorsque l'entité cible à laquelle nous nous intéressons appartient à cette partie, le document aura tendance à être un spam (ou potentiellement non pertinent).
  Pour éliminer ce genre de documents, nous considérons $Sen = \{s_i\}$ l'ensemble des phrases mentionnant l'entité cible E_c dans le document et nous y identifions les entités mentionnées. Notre filtre d'énumération élimine un document si l'une des conditions suivantes est vérifiée :
  - Toutes les phrases s_i appartenant à Sen mentionnent une liste abusive d'entités séparées par une virgule. Formellement, $\forall s_i \in Sen, \exists enum \in s_i$ telle que $enum = E_1, \ldots, E_c, \ldots, E_n$. Nous avons fixé n à 30 en nous basant sur des documents d'apprentissage annotés comme "*garbage*".
  - Le rapport entre le nombre d'entités dans Sen et le nombre total de mots est supérieur à un seuil que nous avons fixé expérimentalement à 0.66. Formellement, soit $r(entités, Sen) = \frac{entités}{|Sen|}$ le rapport du nombre d'entités dans Sen par le nombre total de mots $|Sen|$. Le document est rejeté si $r(entités, Sen) > 0.66$.

- ***Filtre de liens hypertextes*** qui élimine les documents contenant plus de 20 liens ou bien qui mentionnent l'entité uniquement dans une *URL*. Un lien est reconnu par la chaîne de caractères `www` ou `http`.

4.4.2.2 Phase 2 : Mesure des scores de vitalité des documents

Dans cette étape, nous attribuons des scores de vitalité aux documents qui ont passé l'étape 1 selon l'une des deux méthodes que nous avons décrites dans les sections 4.2 et 4.3.

4.4.3 *Baselines* et stratégie d'attribution des scores de confiance

Dans la tâche CCR, un système doit fournir pour chaque couple document-entité, un score de confiance entre 1 et 1000. Pour traduire

4. www.jroller.com/melix/entry/nlp_in_java_a_language

le score de vitalité en score de confiance, nous avons utilisé la stratégie utilisée par Dietz et Dalton (2013) qui affecte un score de confiance 1000 au document ayant le meilleur score, 999 au deuxième, etc.

Pour évaluer l'apport des différentes approches proposées, nous avons utilisé les configurations suivantes comme *baselines*.

- LMD_E : qui attribue un score de vitalité d'un document d par rapport à l'entité E en utilisant l'équation suivante :

$$Score_{Vitalité}(d, E) = \prod_{t \in nom\_canonique(E)} P(t|\theta_d)$$

  telle que la probabilité $P(t|\theta_d)$ est calculée par un modèle de langue avec un lissage de Dirichlet. La collection de référence est constitué des documents d'apprentissage du corpus KBA2013 (datant avant *2012-03-01-00*).
- LMD_ESF : similaire à LMD_E mais sans appliquer les filtres de spams.
- LMD_{PE} : qui attribue un score de vitalité d'un document d par rapport à l'entité E en utilisant l'équation 4.8. Cette équation calcule une certaine similarité entre le modèle du document θ_d et un modèle modèle de pertinence de l'entité θ_{P_E} estimé à partir de la page descriptive P_E. Pour une entité Wikipedia, P_E représente son article Wikipedia (version du 1er Janvier 2012). Pour une entité Twitter, P_E correspond au nom canonique affiché dans son profil Twitter. Pour les topics de 2014, P_E correspond au nom canonique donné dans la définition du topic (figure 4.4).
- *Chrono* : cette *baseline* trie les documents filtrés issus de la première étape par ordre chronologique, c'est à dire le premier document appariant avec l'entité aura un score de confiance 1000, le deuxième aura 999, etc.
- *Booléenne* : cette *baseline* considère les documents filtrés dans la première étape sans aucun tri, c'est à dire que tous les documents ont le même score de confiance.

Dans la section 4.4.4, nous évaluons l'impact des filtres de spams que nous avons définis. La section 4.4.5 compare les différentes façons possibles pour estimer le modèle de vitalité. La section 4.4.6 évalue l'apport de la fraîcheur dans la détection des documents vitaux.

4.4.4 Expérimentations (I) : Impact des filtres de spams

Dans cette section, nous étudions l'effet des filtres de spams utilisés dans la première étape. Nous utilisons les mesures du Rappel et et de Précision calculées comme suit :

$$Rappel = \frac{TP}{P} \tag{4.10}$$

$$Précision = \frac{TP}{TP + FP} \tag{4.11}$$

avec

- P est le nombre de document positifs (spams). Nous considérons les documents *neutres* ou *déchets* comme documents spams.
- TP est le nombre de documents positifs (spams) détectés par un filtre.
- FP est le nombre de documents pertinents (*utiles* ou *vitaux*) faussement rejetés par un filtre. Il s'agit d'une fausse alerte déclenchée par un filtre.

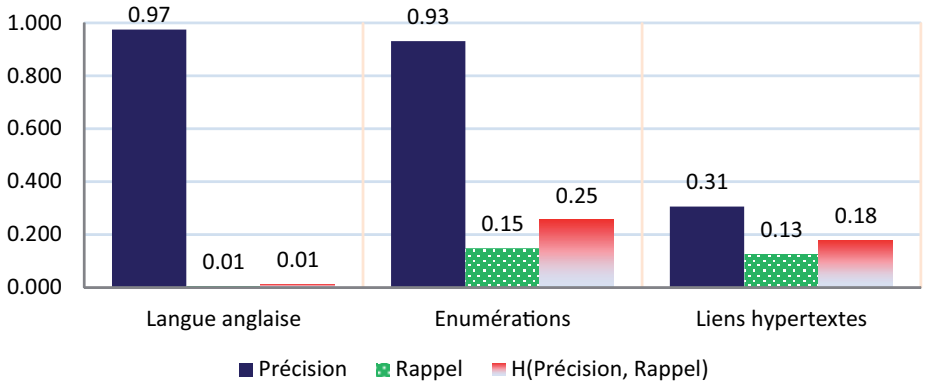

FIGURE 4.6 – Performances des filtres de spams. *H(Précision, Rappel)* est la moyenne harmonique entre le rappel et la précision.

La figure 4.6 montre la performance de chaque filtre en considérant les topics de la tâche CCR 2013. Le filtre de détection des documents anglais élimine uniquement 39 documents non pertinents avec un rappel très faible de 0.01 et une précision de 0.97. Environ 3000 documents sont rejetés par le filtre de liens hypertextes avec un rappel de 0.13. Sa précision est néanmoins faible (0.31). Enfin, le meilleur filtre est celui qui détecte les énumérations avec une bonne précision de 0.93 et un rappel de 0.15 permettant d'éliminer environ 1100 documents spams.

Nous avons évalué l'impact de ces filtres en termes de performance dans le processus de filtrage des documents vitaux. La première ligne du tableau 4.3 considère la configuration LMD_ESF qui n'utilise aucun filtre. Cette configuration obtient une hF1 de 0.358 pour les topics de CCR 2013. Les lignes suivantes montrent l'impact de l'application des filtres de spams

sur LMD_ESF, en considérant la mesure hF1. Les résultats s'améliorent significativement en appliquant tous les filtres (+0.25 en termes de hF1). Ceci prouve **l'importance de l'application des filtres booléens dans la tâche de filtrage des document vitaux**. Cette conclusion a été soulignée également par Efron *et al.* (2014).

TABLE 4.3 – Impact des filtres de spams sur la mesure de hF1 appliqués à la configuration LMD_ESF pour les topics de la tâche CCR 2013

Langue anglaise	***Énumérations***	***Liens hypertextes***	***hF1***
			0.358
		+	0.364
	+		0.372
	+	+	0.379
+			0.361
+		+	0.367
+	+		0.376
+	+	+	**0.383**

4.4.5 Expérimentations (II) : Filtrage des documents vitaux basé sur les modèles de langues de vitalité

4.4.5.1 Configurations du modèle de vitalité

Dans cette section, nous expérimentons différentes configurations afin de répondre aux questions posées dans la section 4.2.

En termes de **généralité** du modèle de vitalité et afin d'évaluer la performance d'un modèle vital général, nous avons considéré deux types de configurations :

- une configuration générale (**Gen**) qui estime un seul modèle de vitalité général s'appliquant à toutes les entités. Ce modèle est estimé en exploitant tous les documents échantillons vitaux disponibles.
- une configuration spécifique (**Spe**) qui estime un modèle de vitalité par entité en exploitant les documents vitaux spécifiques à cette entité.

En termes d'**unité** d'information considérée pour l'estimation de la vitalité et afin de vérifier si une partie de document peut être suffisante pour refléter la vitalité, nous avons considéré différentes unités d'information d'un document (que ce soit un document d'apprentissage ou d'évaluation) :

- tout le contenu du document (**W**, pour *Whole* en anglais),
- la concaténation des paragraphes du document mentionnant une variante de l'entité (**Pg**). Un paragraphe contient 5 phrases au maximum.

- la concaténation des phrases du document mentionnant une variante de l'entité (**Sen**).

Enfin, en termes de **dimension**, nous avons modélisé la vitalité de deux façons différentes :

- de façon unidimensionnelle (**Uni**) telle que décrite dans la section 4.2.1. En prenant ce choix, tous les documents échantillons vitaux auront la même importance.
- de façon multidimensionnelle (**Multi**) telle que décrite dans la section 4.2.2. Nous estimons un modèle vital relatif à chaque document échantillon vital. Le modèle de vitalité final combine et pondère les différents modèles.

Au final, nous expérimentons **12** configurations selon la généralité, l'unité et la dimension considérée. Concernant les paramètres, nous avons fixé expérimentalement le paramètre de lissage μ à 100 et nous avons considéré les 30 premiers termes triés en fonction de leur probabilité dans le modèle de vitalité.

4.4.5.2 Comparaison des différentes configurations basées sur le modèle de vitalité

La figure 4.7 présente les résultats (en termes de hF1) des 12 configurations de notre approche ainsi que la *baseline chrono* pour les entités de la tâche CCR 2013. Nous rappelons que la hF1 correspond à la valeur maximale obtenue parmi tous les points de cutoff. Nous analysons les résultats en termes de généralité, d'unité et de dimension.

- En termes de **généralité**, la première observation importante que nous pouvons effectuer est que l'application d'un modèle de vitalité général (*Gen-**) pour toutes les entités n'améliore pas les résultats par rapport à la *baseline chrono* en particulier lorsqu'on considère les phrases (*Sen*) ou les paragraphes (*Pg*). Cependant, on constate une légère amélioration lorsque tout le contenu du document (*W*) est exploité.
  Par contre, on constate clairement que les résultats obtenus par les configurations spécifiques (*Spe*) dépassent à la fois ceux du modèle général (*Gen-**) et le résultat de *chrono*. Donc, **chaque entité a son propre vocabulaire vital. De plus, l'estimation d'un modèle général est loin d'être performante**.

- En termes d'**unité**, nous remarquons aussi que le fait de ne considérer que les phrases des documents échantillons vitaux *(Sen)* pour l'estimation du modèle vital ne permet d'améliorer que très peu le tri des documents vitaux par rapport à la *baseline chrono*. En

considérant les paragraphes *(Pg)*, les résultats s'améliorent. **Les meilleurs résultats sont obtenus en considérant le document entier** *(W)*.

- En termes de **dimension**, nous remarquons que la modélisation unidimensionnelle de la vitalité *(W-Uni)* semble être légèrement meilleure qu'une modélisation multidimensionnelle *(W-Multi)* lorsqu'on considère tout le contenu des documents *(W)*. Cependant, cette amélioration n'est pas significative. Nous constatons même un résultat inverse lorsqu'on considère les phrases *(Sen)* ou les paragraphes *(Pg)*. Donc, **il n'existe pas de différence significative entre la modélisation multidimensionnelle et la modélisation unidimensionnelle**.

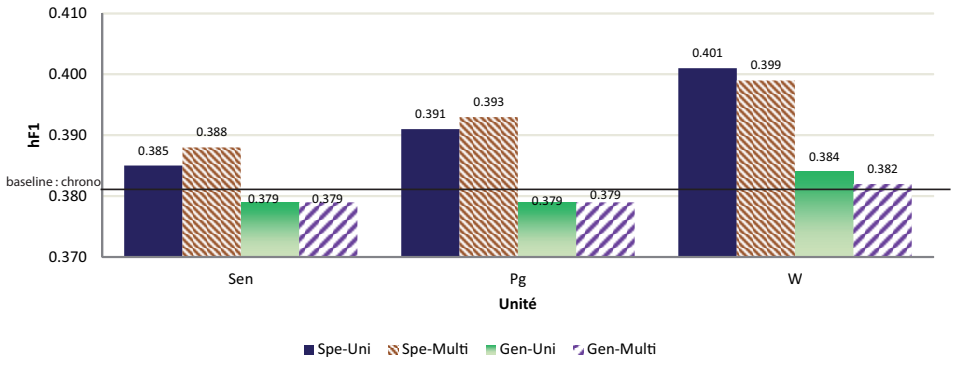

FIGURE 4.7 – Résultats des différentes configurations de notre approche basée sur le modèle de vitalité pour les entités de la tâche CCR 2013

Ensuite, nous illustrons dans la figure 4.8 la performance des configurations en termes de $F_{mesure}@i$ pour les 200 premiers documents sélectionnés. Nous traçons uniquement les résultats des configurations *(*-Uni)* (puisqu'il n'existe pas de différence significative entre *(*-Uni)* et *(*-Multi)*).

Nous remarquons là encore que l'unité d'information choisie *(Sen, Pg, W)* a un impact important sur la performance obtenue ($F_{mesure@i}$) en particulier dans les 150 premiers points de cutoff (>850) avant de converger presque vers la même $F_{mesure@i}$ quand on arrive vers le 200ème document sélectionné. En effet, le fait de considérer les paragraphes *(Pg)* permet de mieux capter la vitalité par rapport à l'exploitation des phrases *(Sen)*. **Les meilleures performances pour tous les points de cutoff sont obtenues en considérant tout le contenu des documents *(W)***.

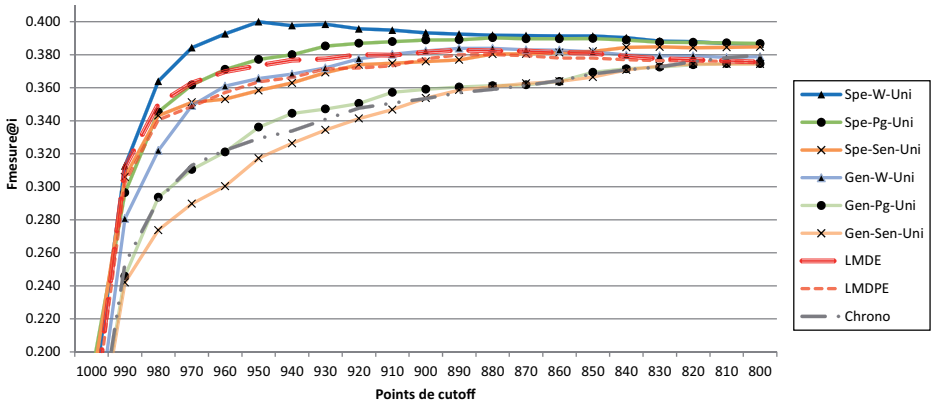

FIGURE 4.8 – Performances de nos configurations (modèles de vitalité) en fonction des points de cutoffs pour les entités de la tâche CCR 2013

Pour comprendre en détails l'impact de l'unité choisie sur les résultats, nous comparons dans la figure 4.9 la différence en hF1 pour chaque topic entre la configuration *(Spe-W-Uni)* et la configuration *(Spe-Pg-Uni)* (barres vertes), et entre la configuration *(Spe-W-Uni)* et la configuration *(Spe-Sen-Uni)* (barres rouges). Nous pouvons voir que pour les 27 premiers topics, l'exploitation de tout le contenu des documents *(Spe-W-Uni)* est meilleure que l'exploitation des paragraphes *(Spe-Pg-Uni)*. Par contre, pour les 17 derniers topics, les résultats sont inverses.

En effet, les mêmes termes vitaux sélectionnés (30 termes dans notre cas) diffèrent en fonction de la configuration (**-Sen*, **-Pg*, ou **-W*). Pour mieux comprendre l'avantage ou l'inconvénient de la considération de tout le contenu des documents **-W* par rapport à l'exploitation unique de *Sen* ou *Pg*, nous avons examiné quelques exemples du corpus KBA 2013.

- Considérer *(W)* est plus avantageux que *(Pg)* ou *(Sen)* pour :
  - l'entité `Fargo-Moorhead Symphony Orchestra`. En effet, la plupart des documents vitaux pour cette entité sont issus du site web `http://www.inforum.com/` (Voir figure 4.10). La configuration *Spe-W-Uni* a réussi à apprendre certains termes relatifs au vocabulaire du site web tels que `news`, `inforum`, `forum`, `communications`, `company`, `like` alors que les configurations *Spe-Sen-Uni* et *Spe-Pg-Uni* ont échoué à capter ces termes.
    Donc, ***(W)* permet de détecter des termes décrivant le vocabulaire typique du site web qui sont utiles pour identifier les futurs documents vitaux dans ce même site web**.

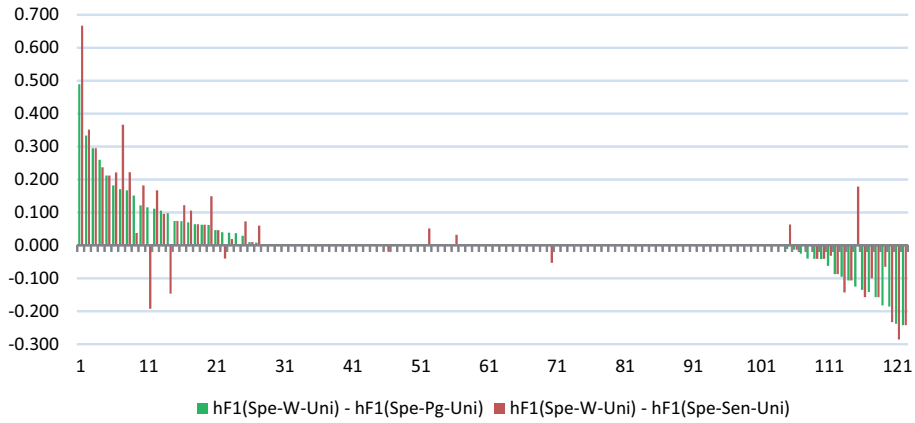

FIGURE 4.9 – Impact de l'unité choisie (W, Pg, Sen) sur la performance de chaque topic de la tâche CCR 2013. Les barres vertes (respectivement rouges) représentent la différence en termes de hF1 entre la configuration Spe-W-Uni et la configuration Spe-Pg-Uni (respectivement Spe-Sen-Uni)

FIGURE 4.10 – Site Web reportant un document vital pour l'entité `Fargo-Moorhead Symphony Orchestra`. Les termes vitaux relatifs au vocabulaire du site web sont encerclés.

- les entités `Atacocha` et `Richard Edlund`. Par exemple, pour l'entité `Atacocha` (Société minière), la configuration *Spe-W-Uni* donne des poids plus importants aux termes `gold`, `silver`, `copper`, `energy`, `increased`, `iron`, `metals` qui n'existent pas ou auront moins d'importance dans les modèles vitaux appris par *Spe-Sen-Uni* et *Spe-Pg-Uni*.
  Donc, **le fait de considérer tout le contenu des documents (**-W*) permet de mieux capter les termes les plus représentatifs ayant un nombre d'occurrences important, ne surtout lorsque le document ne parle que de l'entité.** Ces termes seront moins importants dans *Sen* ou *Pg* car leur fréquence d'apparition est faible.

- Considérer (W) est moins avantageux que *(Pg)* ou *(Sen)* pour :
  - les entités `Paul Johnsgard`, `Toquepala mine` et `Phyllis Lambert`. En effet, les documents vitaux se focalisent sur l'entité considérée uniquement dans une seule section. Par conséquent, le modèle vital appris en considérant (W) peut éventuellement dévier du sujet principal. Donc, **lorsque le document parle de plusieurs sujets, considérer les phrases *(Sen)* ou les paragraphes *(Pg)* semble être plus avantageux que l'exploitation de tout le contenu du document *(W)***.

  - les entités `Fargo Air Museum` et `Jamie Parsley`. En effet, les documents vitaux relatifs à ces entités mentionnent certains termes décrivant le contexte temporel de l'information reportée sur l'entité par exemple `last`, `week`, `today` ... ou bien certains verbes d'action associés à l'entité tels que `visit`, `said`, etc. Ces termes ("temporels" ou "verbes d'action") sont moins fréquents que d'autres termes dans le document, par la suite ils ont plus de chance d'être détectés par *(Pg)* ou *(Sen)* que par *(W)*.

    Donc, **l'exploitation de tout le contenu du document (*W*) manque parfois certains termes "temporels" ou "verbes d'action" qui semblent être utiles pour détecter la vitalité**.

4.4.6 Expérimentations (III) : Filtrage des documents vitaux basé sur les expressions temporelles

4.4.6.1 Configurations basées sur les expressions temporelles

Les expérimentations menées dans cette section ont pour but de prouver l'importance du facteur fraîcheur pour estimer la vitalité des documents. Nous évaluons aussi l'impact de l'unité d'information considérée pour

estimer ce critère.

Le facteur fraîcheur est traduit par le score $Score_{Fraîcheur}$ (équation 4.6). Nous estimons ce score en nous basant sur les dates reconnues à partir des expressions temporelles contenues dans le texte du document (Par exemple les expressions temporelles `Last week`, `At the end of 2012`, `December 31, 2013`, etc.).

Pour identifier et normaliser ces expressions, nous utilisons la bibliothèque SUTIME (Chang et Manning, 2012). Cette bibliothèque utilise la date de publication du document comme date de référence. Par exemple, pour un document publié le `1er janvier 2014 (Wednesday)`, l'expression `Tuesday` est identifiée et normalisée comme `2013-12-31`.

Pour évaluer l'impact de l'unité d'information considérée pour estimer la fraîcheur, nous distinguons les trois configurations suivantes :

- $\mathbf{P\&F_{Sen}}$: La fraîcheur est estimée en considérant les expressions temporelles reconnues dans les phrases mentionnant l'entité.
- $\mathbf{P\&F_{Pg}}$: La fraîcheur est estimée en considérant les expressions temporelles reconnues dans les paragraphes mentionnant l'entité.
- $\mathbf{P\&F_{W}}$: La fraîcheur est estimée en considérant les expressions temporelles reconnues dans tout le contenu du document.

Les configurations ci-dessus combinent deux facteurs : la fraîcheur et la pertinence thématique. Pour montrer l'impact du facteur fraîcheur dans la détection de la vitalité, nous comparons ces configurations avec la *baseline* LMD_{PE} qui exploite uniquement la pertinence thématique.

Notre approche utilise 3 paramètres σ, μ and $top_k(P_E)$. Pour estimer les valeurs optimales de ces paramètres, nous avons utilisé la validation croisée à 3 plis. Nous avons varié $\sigma \in [1, 360]$ (avec un pas de 30), $\mu \in [50, 1000]$ (avec un pas de 50) et $top_k(P_E) \in [5, 30]$ (avec un pas de 5). Les valeurs optimales obtenues sont $\sigma = 30$, $\mu = 200$ et $top_k(P_E) = 20$. Nous avons fixé ϵ de l'équation 4.9 à 10^{-4}.

4.4.6.2 Comparaison des différentes configurations basées sur les expressions temporelles

Le tableau 4.4 compare les différentes configurations de notre approche pour les topics de la tâche CCR 2013. Nous constatons d'abord que la fraîcheur (lignes du tableau ayant le label $P\&F$) améliore la performance par rapport à l'exploitation unique de la pertinence (LMD_{PE}). Cela confirme que **la fraîcheur représente un critère important qui doit être pris en compte pour détecter les documents vitaux**.

Contrairement aux résultats de l'approche basée sur les modèles de

TABLE 4.4 – Comparaison des différentes configurations basées sur les expressions temporelles pour les topics de la tâche CCR 2013. i* correspond au point de cutoff pour lequel $F_{mesure}@i$ est maximale.

	i*	**mPrec.@i***	**mRappel@i***	**hF1**
$P\&F_{Sen}$	**947**	**0.290**	0.591	0.389
$P\&F_{Pg}$	921	0.286	0.660	**0.399**
$P\&F_{W}$	323	0.250	**0.790**	0.379
LMD_{E} (*Baseline*)	223	0.265	0.690	0.383
LMD_{PE} (*Baseline*)	340	0.250	**0.790**	0.379
chrono (*Baseline*)	732	0.254	0.765	0.381

langues (figure 4.7), l'estimation de la fraîcheur en se basant uniquement sur les parties du texte mentionnant l'entité (*Sen* or *Pg*) donne de meilleurs résultats comparativement à la configuration basée sur tout le contenu du document (*W*). Ceci pourrait être expliqué par le fait que considérer des parties du document qui ne sont pas à proximité de l'entité peut apporter des dates fraîches (proches de la date de publication du document) mais qui ne concernent pas l'entité.

Nous constatons par ailleurs que le fait de considérer les expressions temporelles existantes dans les phrases mentionnant l'entité ($P\&F_{Sen}$) dégrade légèrement la performance par rapport à l'exploitation des paragraphes ($P\&F_{Pg}$). Ceci est probablement dû à l'insuffisance du texte considéré dans *Sen* pour identifier toutes les dates relatives à l'entité. Afin de mieux comprendre les résultats et leurs raisons, nous analysons dans la section suivante la présence des dates dans les différentes parties des documents.

4.4.6.3 Présence des dates dans les documents

Dans certains cas, la partie considérée dans le document (*Sen, Pg ou W*) peut ne pas contenir une expression temporelle. La figure 4.11 montre la présence des dates dans les différentes parties des documents comportant les entités étudiées dans tâche CCR 2013. Nous constatons que 93% des documents *(W)* contiennent au moins une expression temporelle. En considérant uniquement les parties mentionnant l'entité, nous pouvons identifier au moins une date dans 76% des paragraphes *(Pg)* et dans 51% des phrases *(Sen)*.

Les figures 4.12, 4.13 et 4.14 tracent en détails la distribution des dates fraîches dans les différentes classes de documents (*vital, utile, non pertinent* respectivement) pour les entités de la tâche CCR 2013, en considérant les différentes parties *(Sen, Pg, W)*.

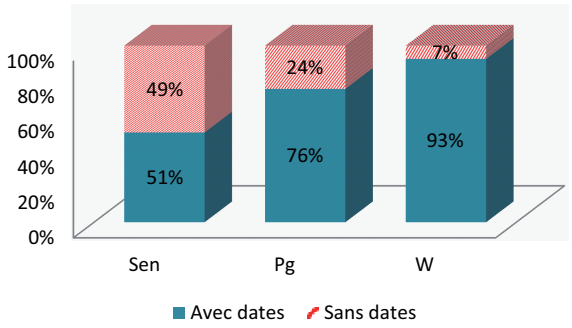

FIGURE 4.11 – Présence des dates dans les différentes parties des documents filtrés pour les entités de la tâche CCR 2013.

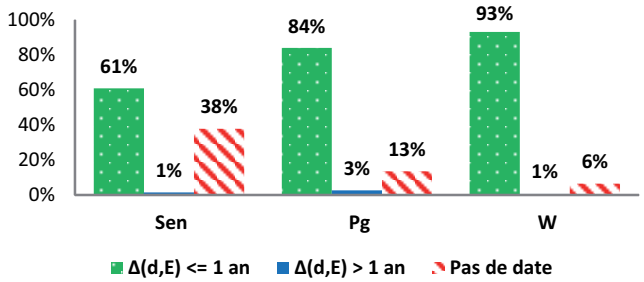

FIGURE 4.12 – Dates identifiées dans les documents *vitaux* pour les entités de la tâche CCR 2013

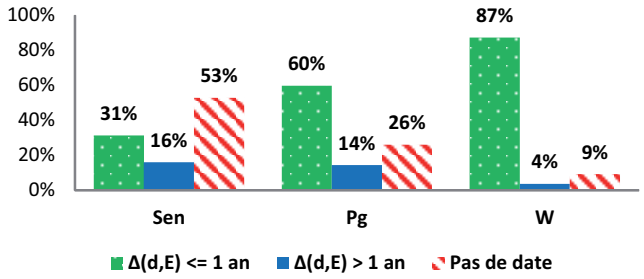

FIGURE 4.13 – Dates identifiées dans les documents *utiles* pour les entités de la tâche CCR 2013

Premièrement, nous remarquons que **les anciennes dates** $(\Delta(d, E) > 1\ an)$ **sont plus susceptibles d'être mentionnées**

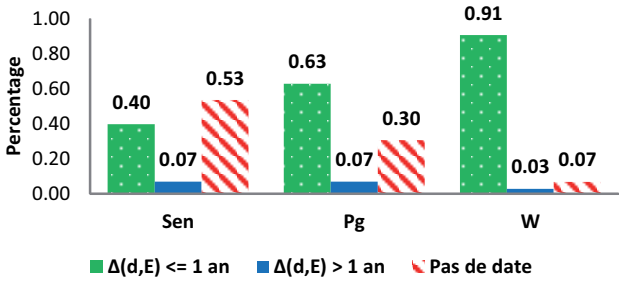

FIGURE 4.14 – Dates identifiées dans les documents *non pertinents* pour les entités de la tâche CCR 2013

dans les documents utiles (figure 4.13) qui décrivent des anciennes informations pertinentes sur l'entité.

Nous constatons également que **le pourcentage des dates fraîches est plus grand dans les documents *vitaux*** (figure 4.12) que dans les documents *utiles* ou *non pertinents* (figures 4.13 et 4.14). La différence est remarquable surtout lorsque nous considérons les phrases ou les paragraphes. En effet,

- **84**% des paragraphes (Pg) de documents vitaux sont "frais" (c'est à dire $\Delta(d.Pg, E) < 1\ an$) contre 60% pour les *Pg utiles* et 63% pour les *Pg non pertinents*.
- **61**% des phrases (Sen) de documents vitaux sont "fraîches" (c'est à dire $\Delta(d.Sen, E) < 1\ an$) contre 31% pour les Sen utiles et 40% pour les *Sen non pertinents*.

  Nous remarquons de plus que 39% des phrases vitales ne contiennent pas de dates fraîches. Alors qu'en considérant les paragraphes, seulement 16% sont non frais. Ceci montre de manière plus claire l'**insuffisance du texte des phrases pour identifier toutes les dates relatives à l'entité**. Ce qui explique la supériorité de la configuration $P\&F_{Pg}$ par rapport à $P\&F_{Sen}$ dans les résultats du tableau 4.4. La faible performance de Sen vient probablement de l'absence des expressions temporelles dans les phrases sélectionnées.
  Une question peut se poser ici : *Est ce que la bonne performance de $P\&F_{Pg}$ est due à la l'importance de cette partie pour détecter la vitalité ou bien à l'insuffisance du texte considéré dans Sen* ?. Autrement dit, supposons que nous avons deux documents. Un premier document $'A'$ mentionnant l'expression *"hier"* (qui fait référence au jour qui précède la date de publication) dans Pg ; un deuxième document $'B'$ mentionnant aussi l'expression *"hier"* uniquement dans Sen. Dans

ces deux documents, le délais optimal $\Delta(d, E)$ est égal à 1 jour. Lequel, parmi ces deux documents, a plus de chance d'être vital ? Pour répondre à cette question nous reportons des analyses supplémentaires dans la section suivante.

4.4.6.4 Probabilité d'un document vital sachant le délai optimal et l'unité d'information considérée

Dans cette section, nous souhaitons déterminer *quelle unité d'information (Sen, Pg, W) détermine mieux la vitalité lorsqu'une date est présente ?* Pour cela, nous analysons les délais optimaux calculés à partir des différentes parties (Sen, Pg, W) des documents contenant les entités étudiées. Nous classons ces délais en intervalles de jours ($]0, 1]$, $]0, 7]$, $]0, 15]$, $]0, 31]$, $]0, 62]$, $]0, 92]$, $]0, 183]$, $]0, 366]$, *et* $]366, +\infty]$) et nous traçons dans la figure 4.15 la probabilité de vitalité d'un document d sachant la classe du délai optimal $\Delta(d, E)$ et l'unité d'information considérée (Sen, Pg ou W).

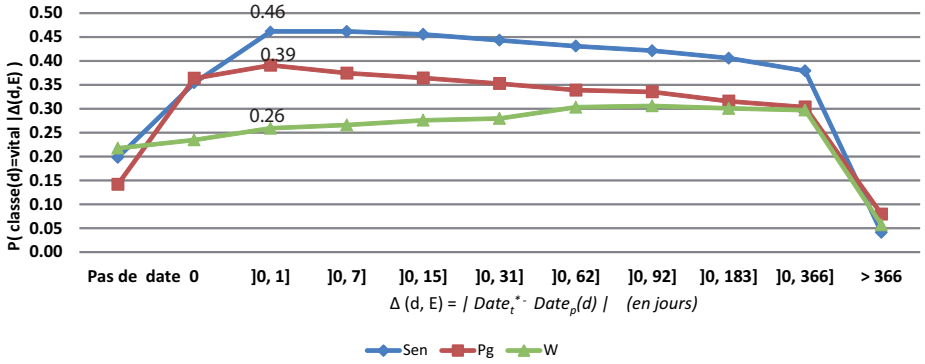

FIGURE 4.15 – Probabilité de vitalité d'un document sachant le délai optimal et l'unité d'information considérée (Sen, Pg, W)

La première remarque importante que nous pouvons soulever est que **la présence d'une date fraîche** ($\Delta(d, E)$<=365 jours) **dans *Sen* indique mieux la vitalité du document que sa présence dans *Pg* ou *W*.** Lorsque le délai $\Delta(d, E)$=1 jour, le cas des documents 'A' et 'B' dans la section précédente, si la partie considérée pour estimer ce délai est *Sen*, alors la probabilité de vitalité du document est de 0.46, sinon si *Pg* est considérée alors la probabilité devient moins faible (0.39). Cette probabilité est encore plus faible (0.26) lorsque *W* est considérée.

D'autre part, **plus le délai** $\Delta(d, E)$ est **minimal**, **meilleure est la probabilité** que le document soit **vital**. Une exception se présente

lorsque $\Delta(d, E) = 0$. Celle ci peut être expliquée par le fait que plusieurs documents non vitaux mentionnent la date de publication dans le document (généralement dans son entête).

Nous remarquons aussi que lorsque $\Delta(d, E)$ est **supérieure à 366 jours, la probabilité de la vitalité d'un document devient très faible**. Cette observation est très importante. Elle peut être exploitée pour rejeter plusieurs document non vitaux.

Dans la section suivante, nous présentons des cas réels extraits à partir du corpus TREC KBA 2013.

4.4.6.5 Cas réels à partir du corpus TREC KBA 2013

Le tableau 4.5 montre quelques extraits de documents du corpus TREC KBA 2013. Dans la première partie, nous montrons des exemples pour lesquels les dates reconnues à partir des documents, soit implicitement (ligne 1) ou bien explicitement (ligne 2), aident à détecter la vitalité (car $\Delta(d, E)$<=1 jour). Le grand délai (265 jours) dans le troisième exemple (ligne 3) permet d'identifier correctement que le document est *utile*.

TABLE 4.5: Cas réels à partir du corpus TREC KBA 2013 montrant l'utilité ou l'insuffisance de l'exploitation des expressions temporelles pour la détection de la vitalité

Ligne	Entité (E)	Phrase (Sen)	$Date_p(d)$	$\Delta(d, E)$	Classe
Cas où les expressions temporelles sont utiles					
1	Atacocha	On ***Friday***, silver miner Minera *Atacocha*, the Lima-based zinc and silver mining company, fell by 5.4%	2012-03-02	1 jour	vital
2	Barbara Liskov	***Monday, 05 March 2012*** *Barbara Liskov* is among the 2012 inductees to the National Inventors Hall of Fame in recognition of her contributions to programming languages and system design.	2012-03-05	0 jour	vital
3	Brenda Weiler	The local chapter's community walk was started in 2006 by *Brenda Weiler*, of Fargo, after she lost her older sister to suicide ***the year prior***.	2012-09-22	265 jours	utile
Cas où les expressions temporelles sont insuffisantes					

4	Angelo Savoldi	The NWA is pleased and proud to induct *Angelo Savoldi* into the Hall of Fame !	2012-11-14	Pas de date !	vital
5	Barbara Liskov	Murray on mathematical biology, *Barbara Liskov* of MIT on in modern programming languages, Ronald Rivest of MIT on cryptography, Leslie G.	2012-04-30	Pas de date !	vital
6	evvnt	The 13th Annual European Shared Services & Outsourcing Week The 13th Annual European Shared Services & Outsourcing Week The 13th Annual European Shared Services & Outsourcing Week offered by *evvnt* will take place in Prague on ***21 May 2013***	2012-11-14	188 jours	vital
7	Bob Bert	Drummer *Bob Bert* played on the album but departed before the tour, and his replacement Steve Shelley remains to ***this day***.	2012-11-08	0 jour	utile
8	Tony Gray	Port continued their preseason schedule with a 4-0 win at Radcliffe Borough on ***Tuesday night*** with goals from Steven Tames (two), Shaun Whalley and *Tony Gray*	2012-08-02	0 jour	non pertinent
9	Alexandra Hamilton	By *Alexandra Hamilton* Email the author ***March 6, 2012*** Tweet Email Print 1 Comment ? Back to Article new Embed | Share	2012-03-06	0 jour	non pertinent
10	Blair Thoreson	*Blair Thoreson* has served in the North Dakota House of Representatives since 1998, representing District 44. Tags : ALEC, American Legislative Economic Council, *Blair Thoreson*, van jones This entry was posted on Wednesday, ***April 18th, 2012 at 12 :19 pm*** and is filed under Blog	2012-04-18	0.5 jour	non pertinent

Dans la deuxième partie du tableau, nous montrons des exemples pour lesquels l'exploitation unique des expressions temporelles est insuffisante pour plusieurs raisons :

- Quelques documents vitaux ne mentionnent pas des dates à proximité de l'entité (lignes 4 et 5),
- Un document pourra être vital même si le délai optimal est grand. C'est le cas dans la ligne 6 où le document reporte un événement à venir dans le futur lointain.
- La normalisation de la date reconnue peut être erronée. C'est le cas dans la ligne 7 où l'expression *'to this day'* réfère à une action dans le passé, alors que l'outil d'identification des dates que nous utilisons a considéré que cette expression réfère au présent.
- Certains documents non pertinents thématiquement peuvent reporter une date fraîche (lignes 8, 9 et 10). Le score de pertinence thématique joue un rôle important dans ce cas. Ce score est mieux estimé lorsqu'on dispose de beaucoup d'informations sur l'entité (généralement pour les entités Wikipedia).

4.4.7 Comparaison de nos approches avec les méthodes proposées dans TREC CCR 2013

Dans cette section, nous comparons nos deux méthodes proposées par rapport aux trois meilleurs systèmes proposés dans la tâche CCR 2013 :

- **BIT-MSRA** (Wang *et al.*, 2013) : Cette méthode utilise un classifieur *Random Forest* avec cinq familles de critères liés à l'entité, au document, à la paire entité-document, des critères temporels et des critères de citations.
- **Umass** (Dietz et Dalton, 2013) : approche de classement (ranking) non supervisée basée sur l'expansion de la requête par des noms d'entités reliées.
- **Udel** (Liu *et al.*, 2013b) : approche de classement (ranking) non supervisée qui exploite le nombre d'occurrences et le poids des entités reliées récupérées à partir des pages Wikipedia.

Nous considérons aussi les deux *baselines* suivantes :

- **TREC-CCR-Baseline** : C'est une *baseline* orientée rappel donnée par les organisateurs de la tâche. Elle sélectionne tous les documents contenant les variantes des entités récupérées manuellement.
- ***Booléenne*** *(phase 1 de notre approche)* : qui sélectionne les documents selon notre méthode présentée l'étape 1 de filtrage (section 4.4.2.1). Les documents sélectionnés ont tous le même score de confiance.

Nous considérons nos meilleures configurations obtenues dans les expérimentations II et III :

- ***Spe-W-Uni*** : qui trie les documents filtrés à l'issue de la première étape à l'aide d'un modèle de vitalité unidimensionnel spécifique à chaque entité.
- $\mathbf{P\&F_{Pg}}$: qui trie les documents filtrés à l'issue de la première étape à l'aide d'un score de vitalité combinant la pertinence et la fraîcheur du document. Le score de fraîcheur est estimé en considérant les expressions temporelles contenues dans les paragraphes mentionnant l'entité.

Nous notons que ces configurations sont évaluées en rejouant la tâche *TREC CCR*, il ne s'agit pas d'une participation officielle.

Tout d'abord, nous comparons les performances des systèmes à l'étape de *filtrage* en considérant tous les documents sélectionnés (i.e., *cutoff=1*). Le tableau 4.6 montre les résultats obtenus.

TABLE 4.6 – Comparaison de nos approches avec les meilleurs systèmes dans la tâche TREC CCR 2013 systèmes sans tenir compte des scores de confiance

	Prec.@1	**Rappel@1**	$\mathbf{F_{mesure}@1}$
Booléenne	**0.249**	**0.790**	**0.379**
BIT-MSRA	0.244	0.650	0.355
TREC-CCR-Baseline	0.190	**0.824**	0.310
Udel	0.199	0.695	0.309
Umass	0.201	0.662	0.309

Les approches non basées sur la classification (*Booléenne, Umass, Udel*) obtiennent de bons résultats au niveau du rappel par rapport à l'approche basée sur la classification (*BIT-MSRA*). Cette dernière est sanctionnée lorsqu'elle n'arrive pas à classer correctement de nombreux documents vitaux.

Le tableau 4.7 compare les différents systèmes en utilisant la mesure officielle de la tâche (hF1). Nos approches (***Spe-W-Uni*** et $\mathbf{P\&F_{Pg}}$) dépassent le meilleur système proposé dans la tâche CCR 2013 (*BIT-MSRA*). Le test de significativité ne peut pas être appliqué pour les mesures $Precision@i^*$, $Rappel@i^*$ et $hF1$ parce que le meilleur point de cutoff i^* et la stratégie d'attribution des scores de confiance sont différents entre les systèmes.

Même si la méthode $P\&F_{Pg}$ est **non supervisée**, elle a pu obtenir un **bon résultat** (hF1) presque égal à la performance de la méthode totalement

TABLE 4.7 – Comparaison de nos approches avec les meilleurs systèmes proposés dans le cadre de la tâche TREC CCR 2013. i* correspond au meilleur point de cutoff.

	i*	**Prec.@i***	**Rappel@i***	**hF1**
Spe-W-UNI	**955**	**0.306**	**0.579**	**0.401**
$P\&F_{Pg}$	921	0.286	0.660	0.399
BIT-MSRA	140	0.257	0.601	0.360
Udel	40	0.199	0.695	0.309
Umass	660	0.216	0.591	0.316

supervisée *Spe-W-Uni*.

TABLE 4.8 – Comparaison de notre approche $P\&F_{Pg}$ avec le meilleur système proposé dans la tâche CCR 2013 : *BIT-MSRA*. † indique une amélioration significative en utilisant le test de *Student* pairé et bilatéral avec p <0.05).

	Precision@T30	**Rappel@T30**	**$\mathbf{F_{mesure}}$@T30**
BIT-MSRA	0.211	0.297	0.170
$P\&F_{Pg}$	**0.286** †	**0.489** †	**0.262** †

Pour évaluer en détails les performances de cette méthode ($P\&F_{Pg}$) par rapport au meilleur système proposé dans la tâche CCR 2013 (*BIT-MSRA*), nous considérons dans le tableau 4.8 le top-30 meilleurs documents selon le score de confiance attribué. Nous avons choisi 30 puisqu'il représente le nombre moyen de documents vitaux par entité. Les résultats montrent que notre approche ($P\&F_{Pg}$) améliore significativement le tri des documents vitaux par rapport au système *BIT-MSRA*.

La figure 4.16 illustre la différence de $F_{mesure}@T30$ entre notre configuration $P\&F_{Pg}$ et le système *BIT-MSRA* pour chacune des entités. Nous constatons que $P\&F_{Pg}$ est meilleure que *BIT-MSRA* pour 69 topics. *BIT-MSRA* quant à lui est meilleur pour 25 topics. Pour mieux comprendre la bonne performance de notre approche par rapport à *BIT-MSRA*, nous considérons l'exemple du topic *Hoboken Volunteer Ambulance Corps*. Cette entité devrait avoir 32 documents vitaux dans la période d'évaluation. En considérant tous les documents sélectionnés indépendamment des scores de confiance (*cutoff=1*), $P\&F_{Pg}$ sélectionne 66 documents dont 31 vitaux, alors que le système de *BIT-MSRA* sélectionne 117 documents dont 29 vitaux. Nous remarquons que les deux méthodes ont un bon taux de rappel (> 90%). Cependant, notre méthode rejette plus de documents non

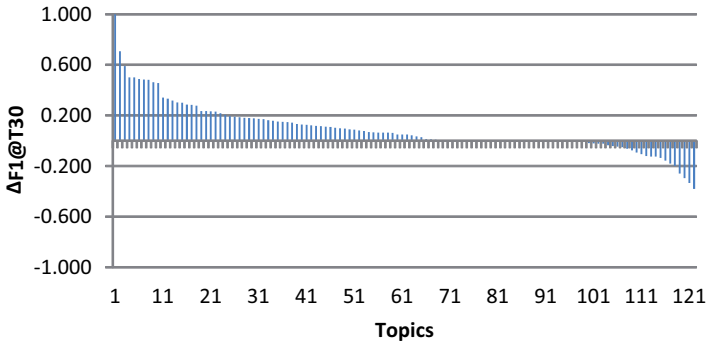

FIGURE 4.16 – Différence en termes de $F_{mesure}@T30$ entre $P\&F_{Pg}$ et *BIT-MSRA* pour chaque topic

pertinents grâce à l'application des filtres de spams (par exemple le troisième document de la figure 4.17). En termes de classement, en considérant le top-30 documents sélectionnés par chaque système, *BIT-MSRA* détecte uniquement un seul document vital, alors que notre configuration ($P\&F_{Pg}$) détecte 16 documents vitaux qui représentent 50% de rappel. Ces documents contiennent des dates fraîches associées à l'entité (Voir les deux premiers documents de la figure 4.17).

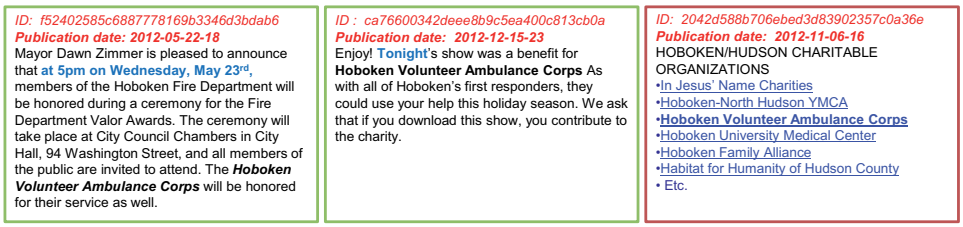

FIGURE 4.17 – Exemples de documents pour le topic *Hoboken Volunteer Ambulance Corps*. Les deux premiers documents sont vitaux et le dernier est non pertinent est rejeté par nos filtres de spams.

4.4.8 Comparaison de nos approches avec les meilleurs méthodes proposées dans TREC CCR 2014

Dans cette section, nous comparons nos deux méthodes proposées par rapport aux meilleurs systèmes proposés dans la tâche CCR 2014 :

- **MSR & KMG** (Jiang *et al.*, 2014) (approche de classification supervisée),
- **uiucGSLIS** (Sherman *et al.*, 2014), précisément nous considérons leur meilleure configuration (supervisée) qui détermine la vitalité d'un

document en calculant sa divergence par rapport à un modèle de langue appris à partir des documents échantillons vitaux de cette entité. Le score de vitalité final est calculé en combinant cette divergence avec d'autres critères a priori estimant la vitalité en se basant sur la longueur et la source du document),

- **BIT_Purdue** (Wang *et al.*, 2014), précisément nous considérons leur meilleure configuration intitulée *GlobalRank* qui est basée sur le *Learning To Rank*),

TABLE 4.9 – Comparaison de nos approches par rapport aux meilleurs systèmes proposés dans le cadre de la tâche TREC CCR 2014.

	i*	**Prec.@i***	**Rappel@i***	**hF1**
MSR & KMG	90	**0.490**	0.737	**0.589**
$P\&F_{Pg}$	759	0.338	0.767	0.469
Spe-W-MULTI	858	0.326	0.829	0.468
uiucGSLIS	24	0.379	0.599	0.465
BIT_Purdue	123	0.301	**0.977**	0.461

Nos deux méthodes obtiennent de bons résultats et se classent en 2ème position parmi les systèmes participant de la tâche CCR 2014. La bonne performance de nos méthodes ainsi que des systèmes *uiucGSLIS* et *BIT_Purdue* **est due principalement au bon taux de rappel**. Cependant, **la précision reste relativement faible** par rapport à la méthode *MSR & KMG*. Cette dernière exploite des critères basés sur les **verbes d'actions qui semblent être très efficaces pour détecter les documents vitaux**. D'autres analyses plus fines devraient être menées afin de mieux comprendre les raisons de cette différence dans la performance.

4.5 Bilan

Dans ce chapitre, nous avons proposé deux méthodes de détection de documents vitaux. La première méthode est supervisée, nécessitant des documents d'apprentissage pour estimer un modèle de vitalité qui servira à détecter les nouveaux documents vitaux. La deuxième méthode est non supervisée. Elle combine la pertinence du document avec sa fraîcheur pour décider de sa vitalité. Nous avons également défini des filtres de spams qui permettent d'éliminer beaucoup de documents non pertinents.

Nous avons évalué nos méthodes dans le cadre de la tâche TREC CCR avec le corpus de 2013 et 2014. Nos systèmes ont obtenu de bons résultats dépassant le meilleur système proposé en 2013.

Les expérimentations menées dans ce chapitre montrent : (1) l'importance de filtres booléens dans l'amélioration de la performance d'un système de détection de documents vitaux ; (2) la vitalité peut être détectée par les modèles de langue spécifiques à chaque entité. L'utilisation d'un seul modèle de langue général ne donne pas de bons résultats ; et (3) la fraîcheur d'un document estimée par les expressions temporelles représente un facteur important pour détecter la vitalité de manière non supervisée.

L'approche basée sur les modèles de langue comporte cependant certaines limites. La première limite réside dans le fait qu'elle nécessite des documents d'apprentissage pour estimer le modèle de vitalité. La deuxième limite concerne l'aspect statique de l'approche. En effet, dans le cas où une entité évolue avec un changement brusque dans le vocabulaire associé (par exemple lorsqu'une personne change de profession, en passant d'acteur à homme politique), la performance du système peut se dégrader.

Enfin, la stratégie d'attribution des scores de confiance pose une autre limite dans nos méthodes utilisées pour répondre à la tâche CCR. En effet, même si l'évaluation de la vitalité est indépendante pour chaque document et n'utilise aucune information future, l'attribution de score de confiance se fait a posteriori. Une approche pour traduire un score de vitalité en score de confiance permettrait de résoudre ce point.

Dans ce chapitre, les approches proposées ne traitent pas le problème de redondance d'information qui ne présentait pas l'objectif de la tâche CCR.

Dans le chapitre suivant, nous nous intéressons à détecter les phrases vitales à partir d'un flux de documents tout en considérant le problème de la redondance d'information.

Chapitre 5

Résumé temporel d'informations sur une entité

5.1 Introduction

Nous présentons dans ce chapitre notre approche d'extraction de phrases vitales et nouvelles concernant une entité. Nous souhaitons construire une synthèse, une sorte de résumé temporel, sur l'entité. Dans la section 5.2, nous présentons l'architecture générale de notre système de filtrage et d'agrégation d'information. La section 5.3 décrit notre approche. Dans la section 5.4, nous présentons les expérimentations menées sur deux collections différentes proposées par la tâche *TREC Temporal Summarization* en 2013 et 2014.

5.2 Architecture générale de notre système de génération de résumé temporel

Nous illustrons dans la figure 5.1 l'architecture générale de notre système de résumé temporel. D'abord, nous filtrons les documents vitaux. Ensuite, nous sélectionnons les phrases vitales et non redondantes.

La *première étape* est importante et nécessite la mise en place d'un processus riche que nous avons détaillé dans le chapitre 4. Dans ce chapitre, nous simplifions cette étape en nous focalisant sur des entités de type événement et en analysant le flux de documents uniquement dans les périodes de déroulement de ces événements. Par conséquent, les documents mentionnant le nom de l'événement ont tendance à reporter des informations vitales.

La *deuxième étape*, à laquelle nous nous intéressons plus particulièrement

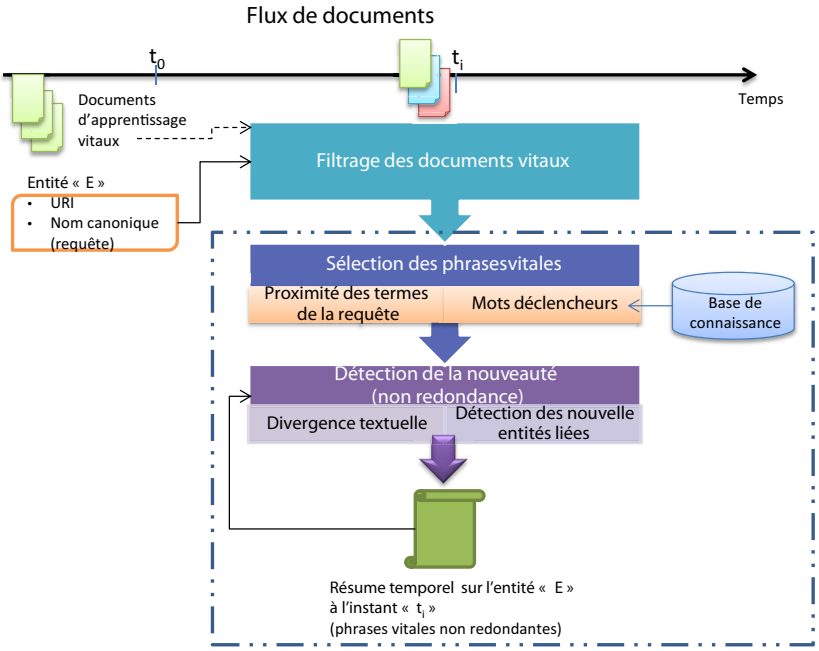

FIGURE 5.1 – Architecture de notre système de résumé temporel

dans ce chapitre, est primordiale puisqu'elle permet de sélectionner les phrases *vitales* et *non redondantes*.

- Concernant la *sélection des phrases vitales*, comme nous l'avons mentionné dans le chapitre 3, plusieurs méthodes se sont basées sur la présence de mots importants. Ces mots sont identifiés soit en exploitant des documents d'apprentissage sélectionnés manuellement (Liu *et al.*, 2013a) soit à partir des documents de flux en utilisant le *Latent Dirichlet Allocation (LDA)* (Chen *et al.*, 2014). Dans notre approche, afin d'identifier les mots importants relatifs à une entité, nous nous basons sur une base de connaissances. Ce genre de sources, malgré sa richesse en informations sur des entités, n'a pas été exploité, à notre connaissance, par les méthodes de l'état de l'art. Notre idée est d'identifier le vocabulaire propre à chaque type d'entités qui sera utilisé pour sélectionner les phrases vitales sur l'entité. Nous exploitons pour cela les informations sur les entités similaires ayant le même type que l'entité considérée dans la base de connaissances. Par exemple, l'entité événement `ouragan Isaac 2012` est similaire à l'entité événement `ouragan Sandy 2012` parce qu'elles sont du même type (`ouragan`).

- Concernant la *détection de la nouveauté (non redondance) des phrases*, les approches de l'état de l'art calculent souvent la divergence textuelle d'une phrase candidate par rapport aux phrases déjà sélectionnées (Xi *et al.*, 2013; Liu *et al.*, 2013a). Pour notre part, nous pensons que l'identification des nouvelles entités dans la phrase candidate peut aider à la détection de sa nouveauté. Pour cela, nous proposons de combiner la divergence textuelle avec l'identification des nouvelles entités liées dans les phrases. Par exemple, l'entité événement `ouragan Isaac` est liée à l'entité `Louisiane` (État du Sud des État-Unis affecté par l'ouragan Isaac en 2012). Lorsqu'une phrase candidate présente une nouvelle entité liée non mentionnée dans les phrases précédentes, elle aura plus de chance d'être nouvelle.

En résumé, nous souhaitons répondre aux trois questions de recherche suivantes :

- A quel point l'exploitation des informations sur les **entités similaires** dans une base de connaissances peut aider à détecter les informations vitales sur l'entité dans un flux de documents Web ?
- Quel est l'apport de la combinaison de la divergence textuelle avec l'identification des nouvelles **entités liées** dans la détection de la redondance dans les phrases vitales ?
- A quel point l'identification des phrases vitales dans un flux de documents peut aider à l'accélération des mises à jour d'une base de connaissances ?

5.3 Approche de génération de résumé temporel basée sur l'exploitation des entités similaires et liées

Notre approche a pour but de détecter en temps réel les phrases reportant les informations vitales relatives à une entité donnée à partir d'un flux de documents issus du Web. Ces phrases doivent être :

- *pertinentes* : concernent l'entité,
- *exhaustives* : couvrent les différentes informations publiées sur l'entité,
- *non redondantes* : ne reportent pas la même information,
- *récentes* : émises sans trop de latence.

Formellement, considérons un flux continu F composé de documents d ayant chacun une date de publication $date_p(d)$ et une séquence de phrases s_j tels que $0 \leq j < l(d)$ où $l(d)$ désigne la longueur du document d en nombre de phrases. Soient $h_0, h_1, \ldots, h_n$ des instants séparés par un intervalle de temps constant (par exemple une heure). Nous désignons par F_{h_i} l'ensemble

de documents du flux tel que $\forall d \in F_{h_i}, h_{i-1} \leq date_p(d) < h_i$.

L'algorithme 1 décrit le fonctionnement général de notre approche de détection des phrases vitales relatives à une entité donnée E. A chaque instant h_i, nous distinguons 2 étapes principales :

1. Détection des documents vitaux $DV_{h_i} \subset F_{h_i}$ par rapport à l'entité E (ligne 2 de l'algorithme 1). Comme nous l'avons mentionné dans l'introduction de ce chapitre, cette étape est simplifiée en supposant que la période durant laquelle les documents ont tendance à être vitaux est donnée. Les méthodes proposées dans le chapitre 4 peuvent aussi être appliquées.
2. Sélection des phrases *vitales* et *nouvelles*.
   - La sélection des phrases vitales se base sur la fonction $est\_vitale(s_j, E)$ de la ligne 5 de l'algorithme 1. Nous détaillons cette étape dans la section 5.3.1.
   - La vérification de la nouveauté des phrases candidates par rapport aux phrases déjà sélectionnées ($\in V(E)$). La détection de la nouveauté est assurée par la fonction $est\_nouvelle(s_j, E)$ de la ligne 5 de l'algorithme 1. Nous détaillons cette étape dans la section 5.3.2.

Algorithme 1 Détection des phrases vitales relatives à une entité

Entrées: F : Un flux de documents
Entrées: E : L'entité considérée identifiée par une *URI* et décrite par une requête Q composée de mots clés.
Entrées: $[h_0, h_n]$: Période d'analyse du flux
Sorties: $V(E) \leftarrow \{\}$: L'ensemble des phrases vitales relatives à E (enrichi à chaque heure h_i) formant le résume temporel de l'entité E.

```
1:  pour chaque i ∈ [1, n] faire
2:    DV_{h_i} ← détection_des_documents_vitaux(F_{h_i}, E)
3:    pour chaque d ∈ DV_{h_i} faire
4:      pour chaque s_j ∈ d faire
5:        si est_vitale(s_j, E)  ET  est_nouvelle(s_j, V(E)) alors
6:          enrichir(V(E), s_j)
7:        fin si
8:      fin pour
9:    fin pour
10: fin pour
```

Dans notre approche, nous analysons le flux de documents en temps quasi réel. La décision de la vitalité d'une phrase doit être prise dans l'heure à laquelle le document est publié, ce qui permet d'émettre des phrases *récentes* à l'utilisateur. Dans notre approche, nous nous intéressons, tout d'abord, à

la *pertinence* et l'*exhaustivité* des phrases (section 5.3.1). Ensuite, nous nous intéressons à la *nouveauté* (*non redondance*) des phrases (section 5.3.2).

5.3.1 Détection de phrases vitales par l'exploitation des entités similaires

Dans cette section, nous analysons les phrases contenues dans les documents vitaux déjà sélectionnés dans la ligne 2 de l'algorithme 1. Pour chaque phrase, nous devons décider si elle est vitale ou pas par rapport à l'entité considérée E.

Notre intuition est de considérer une phrase comme vitale si elle :
- est à **proximité** de l'entité E, c'est à dire à proximité des termes composant la requête Q de l'entité, **et**
- **contient au moins un mot "important"** relativement à l'entité E.

Nous détaillons ces deux conditions dans les deux sous-sections suivantes.

5.3.1.1 Proximité d'une phrase par rapport à une entité

Un document vital par rapport à une entité E mentionne les termes de la requête Q de cette entité dans différentes phrases. Notre but est de sélectionner les phrases qui reportent les nouvelles informations sur l'entité E. Les phrases d'un document vital ne sont pas toutes pertinentes. Pour cela nous favorisons les phrases qui sont à proximité de l'entité E. Nous supposons que plus la phrase se trouve à proximité des termes de la requête, plus elle a la chance d'être pertinente.

Pour estimer la proximité d'une phrase s_j à évaluer par rapport à un terme t de la requête, nous vérifions la présence de ce terme dans la phrase s_j et dans les phrases voisines qui la précèdent $(s_{j-1}, s_{j-2}, ..., s_{j-dmax})$, ou qui la suivent $(s_{j+1}, s_{j+2}, ..., s_{j+dmax})$ telle que $dmax$ est la distance maximale à considérer. Plus les termes de la requête sont proches de (ou contenus dans) la phrase s_j, plus la phrase est pertinente par rapport à l'entité.

Formellement, nous traduisons la pertinence d'une phrase s_j par rapport à l'entité E décrite par la requête Q en un score de proximité calculé selon l'équation suivante :

$$score_{proximité}(s_j, Q) = \frac{1}{|Q|} \sum_{t \in Q} \sum_{k=0}^{dmax} e^{-k \, * \, match(t, s_{j+k}, s_{j-k})} \tag{5.1}$$

- $|Q|$ est le nombre de termes de la requête Q de l'entité E.

- $match(t,\ s_x,\ s_y)$ est égal à 1 si t est contenu dans l'une des phrases s_x et s_y, 0 sinon.
- $dmax$ est la distance maximale à considérer (calculée en nombre de phrases).
- Nous utilisons la fonction exponentielle négative pour décroître l'importance d'un terme t lorsque sa distance par rapport à la phrase s_j augmente.

Nous considérons uniquement les phrases à proximité de l'entité E qui sont proches de l'ensemble des termes composant la requête Q. Pour cela nous définissons un seuil τ_p et nous gardons uniquement les phrases ayant un score de proximité supérieur à ce seuil. La valeur de τ_p peut être déterminée expérimentalement.

La figure 5.2 illustre un exemple de calcul du score de proximité. Nous voulons calculer le score de proximité de la phrase S_1 par rapport à la requête `Q= Hurricane Isaac 2012` décrivant l'entité `https://en.wikipedia.org/wiki/Hurricane_Isaac_(2012)`. Nous fixons $dmax$ à 3 ce qui revient à considérer la phrase S_1 et les 3 phrases voisines (dans ce cas uniquement les phrases qui suivent S_1 car elle représente la première phrase du document). Nous indiquons les termes de l'entité qui apparaissent dans chacune des phrases. Au fur et à mesure que l'on s'éloigne de la phrase S_1, les poids des termes deviennent plus faibles. Par exemple pour une distance $k = 0$, le mot `isaac` a un poids égal à 1, et pour un distance $k = 3$ le poids devient 0.05. Le score de proximité est la somme de tous les poids des termes composant la requête de l'entité, normalisée par le nombre de termes de Q.

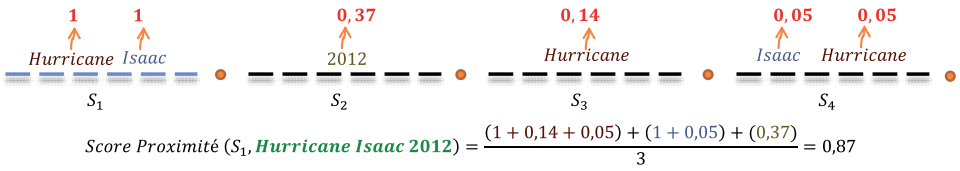

FIGURE 5.2 – Exemple de calcul du score de proximité de la phrase S_1 par rapport à l'entité *Hurricane Isaac 2012*

5.3.1.2 Détection des mots importants à une entité

Dans notre approche, nous supposons qu'une phrase vitale doit contenir un mot important pour l'entité E. Une entité est généralement associée à un ensemble de mots "importants". Lorsqu'un mot important se trouve dans une phrase, il peut refléter sa vitalité. Dans notre approche, nous appelons ces mots des *mots déclencheurs* et nous supposons qu'une phrase

est importante si elle contient au moins un mot déclencheur.

Puisque nous analysons le flux de documents en temps réel, nous ne savons pas a priori quels sont les mots qui peuvent refléter la vitalité d'une phrase. Nous posons l'hypothèse selon laquelle les entités de même type (représentées dans une base de connaissances) partagent les mêmes mots déclencheurs.

Afin d'identifier ces mots déclencheurs, nous proposons d'exploiter toutes les annotations (description en langage naturel) qui ont pu être renseignées sur des entités du type considéré. Nous considérons comme étant une annotation le texte associé à une entité par les propriétés d'annotation de OWL, ou les propriétés du Dublin Core, ou encore le résumé associé dans DBpedia par la propriété dbpedia-owl :abstract.

Par exemple, les mots tels que `effets`, `force`, `tempête`, `blessés`, `dommages` pourront être très utiles pour décrire les entités de type `ouragan` comme ils sont présents dans la plupart des annotations des entités similaires telles que l'`ouragan Sandy 2012` et l'`ouragan Isaac 2012`.

Formellement, soient $X(E) = \{A(E_1), A(E_2), \ldots, A(E_m)\}$ l'ensemble des m extraits des annotations associées aux entités de même type que E. Nous pondérons les termes t par l'équation suivante :

$$\omega(t) = \frac{\sum_{i=1}^{m} TF(t, A(E_i))}{IEF(t)} \tag{5.2}$$

$TF(t, A(E_i))$ est le nombre d'occurrences du terme t dans l'annotation $A(E_i)$

$IEF(t) = log(\frac{m+1}{EF(t)})$ est un facteur utilisé pour donner la priorité aux termes se trouvant dans la plupart des annotations des entités de même type que l'entité E

$EF(t)$ est le nombre d'entités du type dont l'annotation contient le terme t

Les ***top-k*** termes seront considérés comme des mots déclencheurs pour l'entité E.

Dans les expérimentations, nous considérons comme annotations les résumés associés aux entités similaires dans DBpedia par la propriété dbpedia-owl :abstract. Nous étudions l'impact de la sélection autonomique des mots déclencheurs en utilisant l'équation 5.2 par rapport à d'autres méthodes naïves (utilisation des mots clés de la requête comme mots déclencheurs ou bien définir manuellement une liste de mots déclencheurs).

5.3.2 Détection de la nouveauté basée sur l'identification des entités liées

Les phrases sélectionnées à l'étape précédente pourraient contenir des informations vitales redondantes déjà présentes dans le résumé. Afin d'éliminer la redondance, nous comparons chaque phrase vitale candidate à toutes les phrases vitales déjà ajoutées au résumé temporel $V(E)$.

Détecter la nouveauté n'est pas une tâche facile. Comme le montre le tableau 5.1, les deux phrases $s1$ et $s2$ contiennent un grand nombre de mots en commun, mais reportent deux informations différentes. Inversement, les phrases $s2$ et $s4$ sont divergentes textuellement mais portent la même information.

#	Date	Texte
s1	26 Oct. 2012 - 07 :27	**Hurricane Sandy** leaves **21** people dead in **Caribbean**
s2	26 Oct. 2012 - 20 :50	**Hurricane Sandy** leaves **41** people dead in **Caribbean**
s3	28 Oct. 2012 - 08 :08	**Hurricane Sandy** killed at least **52** people in **Haiti** and displaced about **200 000**
s4	30 Oct. 2012 - 06 :17	**Hurricane Sandy** is continuing to head north from the **Caribbean** where it has killed a total of **41** people in the **Caribbean**

TABLE 5.1 – Exemple de phrases vitales

Dans notre approche, nous proposons d'exploiter les entités liées identifiées dans les phrases pour aider à la détermination de la nouveauté/redondance. L'idée est lorsqu'une phrase vitale reporte une nouvelle entité E_i non reportée dans les phrases vitales détectées précédemment, alors elle a tendance à reporter une nouvelle information qui concerne l'entité considérée E et l'entité identifiée E_i.
Par exemple, la phrase $s3$ mentionne une nouvelle entité lieu (`Haïti`) reportant ainsi une nouvelle information vitale (l'ouragan Sandy s'est déplacé vers Haïti).

Nous considérons alors qu'une phrase vitale candidate s_j est nouvelle par rapport aux phrases déjà émises $V(E)$ si son texte est divergent (**DIV**) et/ou présente une entité liée nouvelle (**ELN**) non détectée dans les phrases précédentes $V(E)$.

Formellement, s_j est nouvelle si elle respecte la contrainte de nouveauté suivante :

$$est\_nouvelle(s_j, V(E)) = DIV(s_j, V(E)) \circ ELN(s_j, V(E)) \tag{5.3}$$

$$DIV(s_j, V(E)) = \begin{cases} faux & si\ \exists s_k \in V(E),\ cos(s_j, s_k) > \tau_n(V(E)) \\ vrai & sinon \end{cases} \tag{5.4}$$

$$ELN(s_j, V(E)) = \begin{cases} vrai & si\ \exists x \in EL(s_j, E),\ \forall s_k \in V(E)\ x \notin EL(s_k, E) \\ faux & sinon \end{cases} \tag{5.5}$$

- $EL(s_i, E)$ est l'ensemble des entités liées identifiées dans la phrase s_i.
- $\tau_n(V(E))$ est un seuil de nouveauté textuelle. Au fur et à mesure que l'ensemble de phrases vitales $V(E)$ s'enrichit, le risque de redondance augmente, d'où l'idée de faire décroître le seuil τ_n selon une fonction gaussienne :

  $$\tau_n(V(E)) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{|V(E)|^2}{\delta^2}} \tag{5.6}$$

  Le paramètre σ a un impact sur la tolérance de la divergence, et le paramètre δ contrôle le taux de décroissance du seuil. $|V(E)|$ est le nombre de phrases de l'ensemble $V(E)$.

- Le symbole $\circ$ de l'équation 5.3 peut être un opérateur ***ET*** pour rendre le système orienté Précision en limitant la redondance (dans ce cas, la phrase s4, ne présentant aucune entité liée nouvelle par rapport aux phrases s1 et s2, sera considérée comme redondante malgré que le fait que son texte diverge), ou bien un opérateur ***OU*** pour privilégier le Rappel (dans ce cas, la phrase s2, malgré le fait qu'elle diverge peu par rapport à s1, sera considérée comme nouvelle car elle présente une nouvelle valeur numérique).

Dans les expérimentations, nous étudions l'impact de l'exploitation des entités liées identifiées ainsi que l'opérateur de combinaison (ET/OU) utilisé sur la performance de l'approche.

5.4 Expérimentations

Nous avons évalué notre approche dans le cadre de la tâche *Temporal Summarization* (TS) proposée par la campagne d'évaluation *TREC* en 2013 et 2014. Dans cette section, nous décrivons le cadre expérimental dans lequel nous avons mené nos expérimentations : nous présentons les topics utilisés, les corpus et la manière dont les jugements de pertinence ont été effectués. Ensuite, nous présentons les expérimentations menées afin d'étudier l'impact

de l'exploitation des entités similaires et liées dans la détection des phrases vitales et non redondantes. Nous montrons aussi l'utilité d'un système de résumé temporel pour accélérer la mise à jour de l'encyclopédie Wikipedia. Enfin, nous comparons notre approche par rapport aux méthodes proposées dans la tâche TS 2013 et 2014.

5.4.1 Cadre expérimental

5.4.1.1 Topics évalués dans la tâche TS 2013 et 2014

Les topics proposés dans le cadre de cette tâche correspondent à des entités événements d'actualité tels que des manifestations, des accidents ou des catastrophes naturelles. Le tableau 5.2 illustre les 25 événements [1] proposés par les organisateurs de la tâche en 2013 et 2014. Les colonnes *Début* et *Fin* définissent la période à surveiller pour chaque événement. La période moyenne d'analyse est de 12 jours. Nous notons que le topic 7 a été retiré par les organisateurs à cause de l'absence d'informations pertinentes dans le corpus.

5.4.1.2 Corpus de TREC Temporal Summarization 2013 et 2014

La tâche *Temporal Summarization* (TS) 2013, utilise le même corpus que *KBA 2013* dont les statistiques sont reportées dans le tableau 4.1. Pour la tâche TS 2014, une version filtrée a été élaborée par les organisateurs de la tâche. Cette version contient uniquement les documents s'appariant avec les topics proposés. Pour ce faire, les organisateurs ont créé une requête pour chaque topic (colonne 3 du tableau 5.2). Le top 1000 documents pour chaque requête dans chaque heure sont renvoyés en utilisant le modèle de recherche *BM25* avec les paramètres par défaut définis dans la plate-forme Terrier [2]. La version filtrée élaborée a une taille de *741 GB*.

Nous avons utilisé la bibliothèque open source *Lucene* pour indexer et rechercher des documents. Chaque répertoire (heure) est indexé dans un index à part pour simuler le corpus en tant qu'un flux de documents.

5.4.1.3 Informations vitales et jugements de pertinence

Le but de la tâche Temporal Summarization est de concevoir des systèmes capables de surveiller les événements en détectant à la volée toutes les nouvelles informations publiées dans un flux de documents. Les systèmes doivent extraire les phrases contenant des informations vitales tout en évitant la redondance. Ces informations sont extraites à partir des différentes mises à jour des pages Wikipedia de ces événements.

1. Les *topics* sont disponibles dans ce lien *www.trec-ts.org/documents*.
2. http ://terrier.org/docs/v4.0/javadoc/org/terrier/matching/models/BM25.html

TABLE 5.2 – Topics proposés dans la tâche Temporal Summarization en 2013 et 2014

Topics de TS 2013					
#	**Identifiant (Wikipedia)**	**Requête**	**Type**	**Début**	**Fin**
1	2012_Buenos_Aires rail_ disaster	buenos aires train crash	accident	2012-02-22-12	2012-03-03-11
2	2012_Pakistan_garment_factory_ fires	pakistan factory fire	accident	2012-09-11-13	2012-09-21-13
3	2012_Aurora_shooting	colorado shooting	shooting	2012-07-20-06	2012-07-30-06
4	Wisconsin_Sikh_temple_ shooting	sikh temple shooting	shooting	2012-08-05-15	2012-08-15-15
5	Hurricane_Isaac_(2012)	Hurricane Isaac 2012	storm	2012-08-28-16	2012-09-07-16
6	Hurricane_Sandy	Hurricane Sandy	storm	2012-10-24-15	2012-11-03-15
7	June_2012_North_American_ derecho	June 2012 North American derecho	storm	2012-06-29-15	2012-07-09-15
8	Typhoon_Bopha	typhoon bopha	storm	2012-11-30-14	2012-12-10-14
9	2012_Guatemala_earthquake	guatemala earthquake	earthquake	2012-11-07-16	2012-11-17-16
10	2012_Tel_Aviv_bus_bombing	tel aviv bus bombing	bombing	2012-11-21-10	2012-12-01-10
Topics de TS 2014					
N	**Identifiant (Wikpedia)**	**Requête**	**Type**	**Début**	**Fin**
11	Costa_Concordia_disaster	costa concordia	accident	2012-01-13-21	2013-02-03-00
12	Early_2012_European_ cold_wave	european cold wave	storm	2012-01-22-00	2012-02-18-00
13	Cyclone_Oswald	queensland floods	storm	2013-01-17-00	2013-01-30-00
14	Boston_Marathon_bombings	boston marathon bombing	bombing	2013-04-15-18	2013-04-20-23
15	Port_Said_Stadium_riot	egyptian riots	riot	2012-02-01-13	2012-02-11-13
16	2012_Afghanistan_Quran_burning _protests	quran burning protests	protest	2012-02-20-17	2012-02-28-00
17	In_Amenas_hostage_crisis	in amenas hostage crisis	hostage	2013-01-16-00	2013-01-20-00
18	2011-13_Russian_protests	russian protests	protest	2011-12-04-00	2011-12-25-00
19	2012_Romanian_protests	romanian protests	protest	2012-01-12-00	2012-01-26-00
20	2012-13_Egyptian_protests	egyptian protests	protest	2012-11-18-00	2012-12-01-00
21	Chelyabinsk_meteor	russia meteor	impact event	2013-02-15-03	2013-02-25-03
22	2013_Bulgarian_protests_against_ the_Borisov_cabinet	bulgarian protests	protest	2013-02-10-00	2013-02-20-23
23	2013_Shahbag_protests	shahbag protests	protest	2013-02-05-00	2013-02-22-23
24	February_2013_nor'easter	nor'easter	storm	2013-02-07-00	2013-02-18-23
25	Christopher_Dorner_shootings _and_manhunt	Southern California shooting	shooting	2013-02-03-00	2013-02-13-07

Une phrase est jugée pertinente si elle peut être associée à au moins une information vitale. Cette association est réalisée par les juges de la tâche. Dans l'exemple du topic `2012_Buenos_Aires_rail_disaster`, la phrase `49 dead, over 500 wounded in Buenos Aires!`, émise le `23-02-2012; 03:21`, est associée à trois informations vitales : `"train accident in Buenos Aires, Argentina"`, `"550 injured"` et `"49 confirmed deaths"`.

Enrichissement de la vérité terrain

La vérité terrain (*qrels*) de la tâche consiste en des associations entre les informations vitales extraites de Wikipedia et les phrases reportant ces informations dans le corpus. Chaque phrase du corpus a un identifiant sous forme *timestamp-docID-position* telle que *timestamp* représente la date de publication du document mesuré en secondes depuis l'époque 1970, *docId* représente l'identifiant du document et *position* représente la position (rang) de la phrase dans le document. La vérité terrain dans la tâche TS 2013 comporte environ 9113 phrases alors que les qrels de TS 2014 contiennent 14651 phrases.

Vu que notre système peut renvoyer une phrase avec un nouveau identifiant ne se trouvant pas dans l'ensemble des identifiants des phrases jugées, nous avons enrichi les jugements de pertinences de la façon suivante : si le texte de cette phrase apparie exactement avec une phrase déjà jugée alors elle sera associée aux mêmes informations vitales de la phrase ayant le même texte, sinon (nous n'avons pas un appariement exact) si la similarité est forte ($cos \in [0.8, 1[$) nous avons demandé aux juges membres de notre équipe d'associer manuellement cette nouvelle phrase aux informations vitales. Les vérités terrains enrichies contiennent 21637 phrases pour TS 2013 et 19645 phrases pour TS 2014.

5.4.1.4 Métriques d'évaluation

Nous utilisons les mesures classiques de Rappel, Précision et leur moyenne harmonique :

$$Rappel = \frac{Nombre\ d'informations\ vitales\ détectées}{Nombre\ total\ d'informations\ vitales} \tag{5.7}$$

$$Précision = \frac{Nombre\ de\ phrases\ vitales}{Nombre\ total\ de\ phrases\ émises} \tag{5.8}$$

$$H = 2 * \frac{Précision * Rappel}{Précision + Rappel} \tag{5.9}$$

Nous utilisons aussi une variante de la précision pour pénaliser la redondance d'informations :

$$Précision_N = \frac{Nombre\ d'informations\ vitales\ détectées}{Nombre\ total\ de\ phrases\ sélectionnées} \tag{5.10}$$

$$H_N = 2 * \frac{Précision_N * Rappel}{Précision_N + Rappel} \tag{5.11}$$

Nous utilisons le rappel, la précision et (5.9) pour mesurer la capacité d'un système à renvoyer les phrases vitales, sans pénaliser la redondance. Pour considérer la nouveauté (pénaliser la redondance), nous utilisons le rappel et les équations 5.10 et 5.11.

5.4.2 Configuration de notre système

Dans la première étape de **détection de documents vitaux**, nous appliquons à chaque **heure** un modèle de langue pour récupérer le **topH** documents dont le titre et le contenu doit apparier avec la requête de l'événement (colonne 3 du tableau 5.2). Le score attribué à chaque document d par rapport à la requête Q de l'entité E est calculé en utilisant l'équation suivante :

$$Score_{Vitalité}(d, E) = \prod_{t \in Q} P(t|\theta_d)$$

telle que la probabilité $P(t|\theta_d)$ est calculée par un modèle de langue avec un lissage de Dirichlet. Nous utilisons les documents publiés dans la même heure que le document d comme collection de référence.

Pour l'étape de ***sélection des phrases vitales candidates***, nous appliquons la méthode expliquée dans la section 5.3.2 qui repose sur la détection du ***topK*** mots déclencheurs décrivant des **entités similaires** à l'entité considérée (Eq. 5.2). Nous désignons cette stratégie par ***ES***.
Afin d'évaluer l'intérêt de l'exploitation des entités similaires ***ES*** pour la détection des mots déclencheurs, nous avons considéré deux autres stratégies simples :

- ***Manuelle*** : Dans cette stratégie, nous avons sélectionné une liste prédéfinie de mots déclencheurs qui apparaissent dans la plupart des pages Wikipedia des événements étudiés. Les mots de cette liste sont : `dead(s)`, `death(s)`, `died`, `kill(s)`, `killed`, `killing`, `injured`, `injuries`, `injuring`, `damage`, `victim(s)`, `survivor(s)`, `wound(s)`, `wounded`.
- ***TermesQ*** : Dans cette stratégie, nous considérons les termes de la requête Q de l'entité comme mots déclencheurs.

Pour la ***détection de la nouveauté***, afin d'évaluer l'intérêt de l'exploitation des entités liées ***EL*** identifiées dans les phrases, nous évaluons les méthodes suivantes :

- ***DIV*** : utilisant uniquement la nouveauté textuelle (Eq. 5.4)
- ***EL*** : utilisant uniquement la reconnaissance d'entités liées (Eq. 5.5).
- ***EL*DIV*** : utilisant la fonction de nouveauté combinée avec un opérateur ET (Eq. 5.3)
- ***EL+DIV*** : utilisant la fonction de nouveauté combinée avec un opérateur OU (Eq. 5.3)
- ***Sans*** : Sans appliquer la nouveauté (émettre toutes les phrases sélectionnées à l'étape 2)

Paramétrage de notre approche
Nous avons appliqué la validation croisée afin de fixer les différents paramètres de notre approche, en faisant varier le nombre de documents sélectionnés par heure entre 1 et 20 avec un pas de 1, *top-k* entre 4 et 40 avec un pas de 2, τ_p entre 0.4 et 1 avec un pas de 0.1, δ entre 10 et 300 avec un pas de 10, σ entre 0.5 et 1 avec un pas de 0.1. Les valeurs optimales obtenues sont : *top-h=10* et *top-k=15*, $\tau_p = 0.8$, $\delta = 200$ et $\sigma = 0.5$.

5.4.3 Résultats

5.4.3.1 Rappel maximum à l'issue de l'étape 1 : Détection des documents vitaux

La figure 5.3 illustre le nombre d'informations vitales à retrouver pour les topics proposés dans la tâche Temporal Summarization en 2013 et 2014. En moyenne, il existe 55 informations vitales à retrouver (en bleu) par entité. Dans la première étape, pour chaque topic, notre système sélectionne les 10 meilleurs documents par heure. Nous avons obtenu 20800 documents pour les 24 topics (soit 866 documents par topic) permettant d'atteindre un rappel maximal d'information vitales de 0.65.

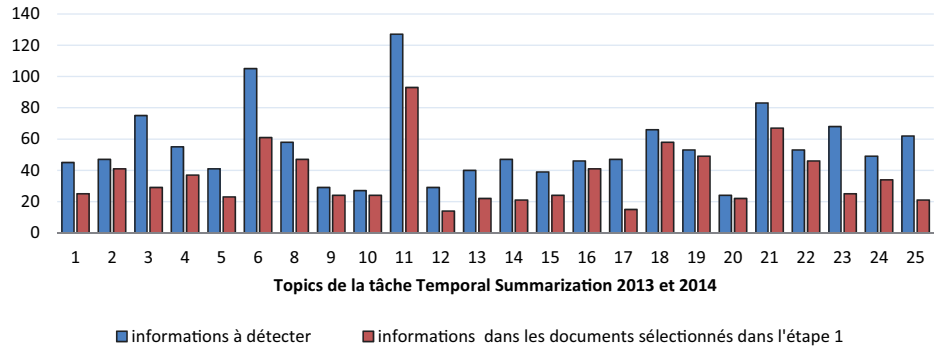

FIGURE 5.3 – Nombre d'informations vitales détectées par topic dans la tâche Temporal Summarization en 2013 et 2014

5.4.3.2 Intérêt de l'exploitation des entités similaires pour la sélection des phrases vitales

La figure 5.4 compare les différentes stratégies de sélection des mots déclencheurs pour la détection des phrases vitales sans tenir compte de la redondance (***sans***). Considérer comme mots déclencheurs uniquement les termes de la requête de l'entité (***TermesQ***) permet de capturer environ **63%** (0.413/0.650) des informations vitales contenues dans les documents sélectionnés dans la première étape avec une précision ne dépassant pas ***0.161***. La condition de proximité (Eq. 5.1) avec un seuil $\tau_p = 0.8$ semble être stricte car elle exige la présence de la plupart des termes de la requête dans les phrases vitales ce qui peut expliquer la perte de 37% d'informations vitales. L'utilisation de la stratégie **ES** revient à vérifier la présence simultanée des termes de la requête et d'un mot déclencheur. Comme résultat, on constate une amélioration légère de la précision par rapport à TermesQ "pratiquement" sans baisse du rappel. Cette stabilité

du rappel prouve que **les mots saillants récupérés automatiquement des annotations associées aux entités du même type permettent de couvrir les différents aspects de l'entité traitée**. L'amélioration de la précision montre l'**importance de ces mots**. La sélection manuelle des mots déclencheurs (**Manuelle**) améliore la précision (surtout pour les entités de TS 2014) mais le rappel est relativement faible par rapport à notre méthode automatique *ES*.

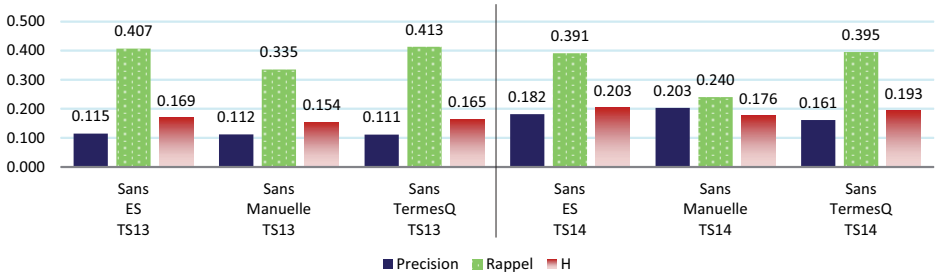

FIGURE 5.4 – Comparaison des différentes stratégies de sélection des mots déclencheurs (ES, Manuelle, TermesQ) pour la détection des phrases vitales

5.4.3.3 Intérêt de l'exploitation des entités liées pour la détection de la nouveauté

La figure 5.5 compare les différentes configurations de détection de la nouveauté. L'application du module de nouveauté permet d'améliorer la $precision_N$ en pénalisant le rappel. **Combiner la divergence textuelle avec l'identification d'entités liées *EL*DIV* donne une meilleure moyenne harmonique** H_N entre le rappel et la précision pour les topics de 2013 et 2014. Utiliser la stratégie ***EL+DIV*** reste utile si nous privilégions l'exhaustivité d'informations.

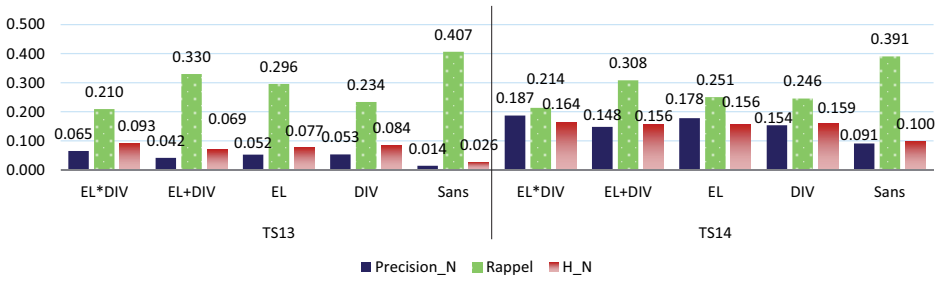

FIGURE 5.5 – Comparaison des différentes méthodes de détection de la nouveauté

5.4.3.4 Comparaison de notre approche par rapport aux participants de tâche Temporal Summarization

Dans cette section, nous comparons notre approche par rapport aux meilleurs systèmes ayant participé à la tâche Temporal Summarization en 2013 et 2014. Nous considérons les systèmes suivants :

- **ICTNET** (Liu *et al.*, 2013a) : Cette approche sélectionne uniquement les documents dont le titre couvre tous les mots clés de la requête. A partir de ces documents, les auteurs ont identifié manuellement une liste de mots déclencheurs pour sélectionner les phrases vitales. Pour détecter la redondance, l'approche applique un algorithme de similarité (*simHash*) qui donne des poids importants aux valeurs numériques.
- **PRIS** : (Zhang *et al.*, 2013) : Cette approche sélectionne les phrases vitales en se basant sur la présence des mots déclencheurs. Ces mots sont identifiés automatiquement à partir des documents de flux en utilisant l'algorithme *Latent Dirichlet Allocation (LDA)*.
- **HLTCOE** (Xu *et al.*, 2013a) : Cette approche se base sur un classifieur pour détecter les phrases pertinentes et contenant de nouvelles informations. Elle utilise des critères mesurant la pertinence du document et de la phrase ainsi que d'autres critères qui exploitent les poids de certains termes représentant des noms d'entités, les prédicats (verbes) et les valeurs numériques.
- **cunlp** (Kedzie *et al.*, 2014) : Cette méthode utilise un classifieur pour déterminer l'importance des phrases en se basant sur des critères sémantiques, géographiques et temporels. Les phrases prédites comme pertinentes par le classifieur sont ensuite regroupées en utilisant un algorithme de propagation d'affinité (Dueck et Frey, 2007). Chaque groupe est supposé contenir des phrases reportant des informations redondantes. Seules les phrases centrales (centres des clusters) seront ajoutées dans le résumé temporel.
- **BJUT** (Zhao *et al.*, 2014) : Cette méthode utilise le modèle BM25 pour sélectionner les documents. Les phrases des documents sont regroupées à posteriori en utilisant l'algorithme de clustering introduit par Zhang *et al.* (1996). Pour chaque événement les auteurs sélectionnent le top-100 phrases représentant les centres des clusters.
- **uogTr** (McCreadie *et al.*, 2014a) : Cette méthode utilise un premier classifieur pour sélectionner les documents qui portent sur le sujet de l'événement et un deuxième classifieur pour sélectionner les phrases vitales en se basant sur un dictionnaire de termes d'urgence, des critères de pertinence de la phrase (par exemple TF-IDF basée sur un corpus de Wikipedia), et des critères de qualité (ex. termes en majuscules). La détection des phrases redondantes est faite par la fonction de similarité de cosinus.

Les résultats sont générés en utilisant l'outil d'évaluation [3] avec les jugements de pertinence officiels. Les mesures *ELG* et *LC* sont similaires aux mesures de précision et rappel respectivement mais en pénalisant la redondance et la latence lors de la détection d'informations (Aslam *et al.*, 2013). Ces mesures sont détaillées dans la section 2.3.2.3.

Notre système aurait pu être classé **premier** (/7 participants) dans la tâche de TS 2013, et **troisième** (/6 participants) pour l'année 2014.

TS 2013				**TS 2014**			
Système	**ELG**	**LC**	**H-ts**	**Système**	**ELG**	**LC**	**H-ts**
*ES ; EL*DIV*	0.1102	0.1986	**0.1355**	cunlp	0.0631	0.3220	**0.1162**
ES ; EL+DIV	0.0768	0.2619	0.1188	BJUT	0.0657	0.4088	0.1110
ICTNET	0.0794	0.3636	0.1078	*ES ; EL*DIV*	0.0881	0.1646	0.1047
PRIS	0.1360	0.1950	0.1029	uogTr	0.0467	0.4453	0.0986
HLTCOE	0.0522	0.2834	0.0827	*ES ; EL+DIV*	0.0712	0.2181	0.0963

TABLE 5.3 – Comparaison de notre système par rapport aux systèmes participants à la tâche TS 2013 et 2014. H-ts est la moyenne harmonique entre ELG et LC.

5.4.3.5 Intérêt de notre approche pour accélérer la mise à jour d'une base de connaissances

Dans cette section, nous prenons Wikipedia comme un exemple de base de connaissances à mettre à jour. Pour étudier l'intérêt de notre approche pour accélérer la mise à jour de Wikipedia, nous analysons la différence entre le temps de détection des informations vitales dans les corpus (issus du Web) et le temps des mises à jour de ces informations dans Wikipedia.

La figure 5.6 compare la rapidité de notre système (ES ; EL*DIV) à détecter les informations vitales pour les 24 événements par rapport aux mises à jour de Wikipedia. Notre système permet de détecter 67% d'informations vitales avant que celles-ci soient mises à jour dans Wikipedia. La moitié des informations sont détectées par notre système 7 heures (au moins) avant qu'elles ne soient mises à jour dans Wikipedia. En moyenne, notre système permet de gagner 18 heures.

Dans le tableau 5.4, nous illustrons quelques exemples d'informations vitales détectées par notre système avant qu'elles soient ajoutées dans Wikipedia.

La figure 5.6 et les exemples du tableau 5.4 montrent l'**intérêt d'exploiter les documents Web pour accélérer la mise à jour des bases de connaissances**. En effet, les informations sont généralement

3. http ://www.trec-ts.org/documents

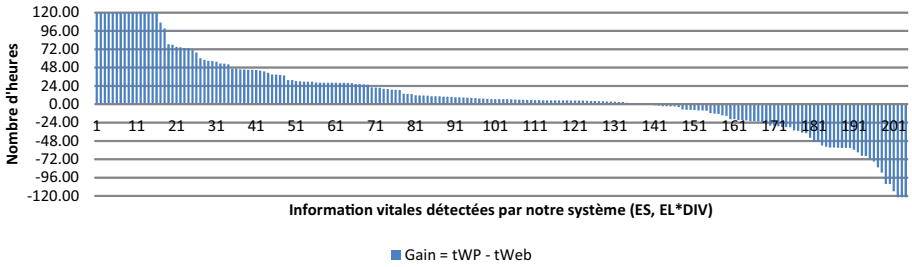

FIGURE 5.6 – Évaluation de la rapidité de notre système (ES ; EL*DIV) par rapport aux mises à jours Wikipedia

Topic	***Information vitale détectée***	**tWeb**	**tWP**	**tIB**	***Gain***
1	*550 injured*	22-02-12 16 :05	22-02-12 22 :49	22-02-12 22 :49	6.7h
1	*crashed at speed of 26 kilometers per hour*	22-02-12 22 :21	22-02-12 23 :01	-	0.67h
9	*39 casualties reported in Guatamala*	08-11-12 00 :33	08-11-12 04 :33	08-11-12 04 :33	1h
9	*48 casualties reported*	08-11-12 07 :42	08-11-12 07 :55	08-11-12 07 :55	0.22h
19	*Early modest estimates put over 5000 people in the streets of Romanian cities*	16-01-12 03 :58	18-01-12 02 :28	-	46.5h
19	*Queensland floods*	27-01-13 11 :35	24-01-13 22 :42	-	60.8h

TABLE 5.4 – Exemple d'informations vitales détectées par notre approche (ES ; EL*DIV). $tWeb$, tWP, tIB représentent les temps de la disposition de l'information par notre système, dans Wikipedia et dans les infoboxes de Wikipedia respectivement. - indique que l'information n'est pas disponible. $Gain = tWP - tWeb$ représente le temps gagné par notre système par rapport à la date de mise à jour de la page Wikipedia de l'événement.

publiées dans le documents Web (presses, blogs, etc.) avant qu'elles soient éditées dans Wikipedia. Bien que les entités analysées représentent des événements largement connus, qui intéressent plusieurs contributeurs, nous constatons qu'il y a toujours un temps de latence. Ce temps de latence devrait être plus grand pour des entités "moins populaires".

Construire un résumé temporel est très important aussi surtout pour les entités populaires pour lesquelles un grand volume de documents sont publiés dans le Web. Ce résumé vise à détecter uniquement les phrases nécessaires pour mettre jour les bases de connaissances tout en évitant la redondance. Notre configuration *ES; EL*DIV* détecte en moyenne environ 140 phrases par entité. La configuration *ES; EL+DIV* favorise l'exhaustivité en renvoyant en moyenne 612 phrases par entité.

Nous notons aussi que la mise à jour n'est pas forcément reportée dans les InfoBoxes principalement exploités pour enrichir les bases de connaissances structurées telles que DBpedia.

5.5 Bilan

Dans ce chapitre, nous avons présenté une approche d'extraction et d'agrégation des phrases vitales issues d'un flux de documents. Nous avons présenté notre approche de résumé temporel qui se base sur l'exploitation des entités similaires et liées à l'entité traitée. Les entités similaires sont obtenues à partir d'une ressource externe (dans ce travail, il s'agit de DBpedia) et permettent d'identifier le vocabulaire propre au type de l'entité considérée. Ce vocabulaire permet d'identifier les phrases vitales à partir des documents. Afin d'éviter de sélectionner des phrases vitales redondantes, nous avons proposé une fonction de nouveauté qui combine la divergence textuelle et les entités liées reconnues dans les phrases.

Les expérimentations menées ont montré que l'exploitation des annotations associées aux entités similaires est utile pour capturer les différents aspects décrivant l'entité. Nous avons aussi montré l'intérêt d'exploiter des entités liées dans la détection de la nouveauté. Finalement, nous avons montré que l'établissement d'un résumé temporel permet d'accélérer la mise à jour des articles Wikipedia. Ceci aura plus d'intérêt lorsqu'il s'agit d'entités moins populaires.

Chapitre 6

Conclusion

6.1 Synthèse des propositions

Les travaux de cette thèse s'inscrivent dans le cadre du filtrage et d'agrégation d'informations vitales relatives à des entités, mais peuvent être appliqués pour faciliter la mise en place des techniques d'ingénierie de connaissances en permettant la sélection des textes à considérer en amont de la phase d'extraction de connaissances. Ce type de tâche est souvent mis en place selon deux étapes, la première porte sur la détection de documents vitaux et la seconde concerne la génération d'une synthèse, une sorte de résumé temporel.

Nos contributions ont porté sur les deux étapes :

- Dans la première étape, nous avons d'abord proposé des *filtres booléens visant à éliminer les documents spams* du flux. Ensuite nous avons proposé deux méthodes pour *évaluer la vitalité des documents filtrés.*
  - La *première méthode* est *supervisée*, elle évalue la vitalité d'un document en exploitant les modèles de langue, plus précisément les modèles de pertinence. Nous pensons que, comme pour la notion de pertinence, il existe un vocabulaire (un modèle de langue) permettant de représenter la notion de vitalité. Nous avons en effet adapté le modèle de pertinence pour estimer la notion de vitalité. Nous avons montré que chaque entité a son propre vocabulaire vital. L'estimation d'un seul modèle de vitalité global à appliquer pour toutes les entités n'a pas d'intérêt. Nous avons aussi montré que la vitalité est mieux estimée en considérant tout le contenu du document qu'en ne considérant que des parties (les phrases ou les paragraphes) mentionnant l'entité. Considérer tout le contenu d'un document permet en effet de capter les termes les plus représentatifs ayant un nombre d'occurrence important ou encore certains termes relatifs au vocabulaire typique du site Web de l'entité.

- La *seconde méthode* est *non supervisée*, elle est basée sur la notion de fraîcheur du document, estimée en considérant les expressions temporelles mentionnées dans son contenu. Nous pensons que la présence d'une expression temporelle, mentionnée à proximité de l'entité, et faisant référence à une date fraîche par rapport à la date de publication du document, peut inférer la vitalité du document. Les expérimentations menées ont montré l'importance de la fraîcheur dans la détection de la vitalité. Des analyses plus détaillées ont montré que la présence d'une date fraîche dans une phrase mentionnant l'entité représente un bon indicateur de vitalité.

- Dans la deuxième étape, nous avons proposé deux méthodes pour résoudre les problématiques relatives à la *sélection des phrases vitales* et à la *détection de la nouveauté (non redondance)* :
  - Pour la *sélection des phrases vitales*, nous avons proposé une méthode qui extrait uniquement les phrases vitales reportant des informations nouvelles sur l'entité à partir des documents vitaux détectés dans l'étape précédente. Nous considérons comme informations nouvelles, toute phrase qui comporte des mots décrivant ou caractérisant l'entité. Pour détecter ces mots, nous avons exploité les annotations associées aux entités similaires (ayant le même type que l'entité considérée) dans *DBpedia*. Les expérimentations ont montré l'importance des mots déclencheurs, qui permettent d'améliorer la précision des phrases tout en couvrant les différents aspects de l'entité (stabilité dans le rappel).

  - Pour la *détection de la nouveauté*, nous avons proposé une méthode qui combine le facteur de divergence textuelle souvent utilisé dans la littérature, avec un deuxième facteur basé sur l'identification des nouvelles entités dans les phrases. Les résultats des expérimentations ont montré que la combinaison conjonctive de ces deux facteurs permet une légère amélioration en termes de précision et de moyenne harmonique.

Pour réaliser nos expérimentations, nous nous sommes basés sur les corpus de documents fournis par la campagne d'évaluation TREC dans les tâches *Knowledge Base Acceleration* et *Temporal Summarization* des années 2013 et 2014.

6.2 Perspectives

Les perspectives de nos travaux portent sur plusieurs points :

- Tout d'abord, à très court terme, nous pensons reprendre les expérimentations pour mieux appréhender le facteur temporel. Plus précisément, dans l'évaluation des méthodes de filtrage, nous avons utilisé la précision, le rappel ainsi que la mesure hF1 (mesure officielle de la tâche) pour évaluer les différents systèmes. Cependant, cette évaluation mesure la performance d'un système en moyenne et non pas dans des périodes de temps différentes. Par conséquent, nous ne pouvons pas analyser si la performance d'un système est stable ou bien si elle se dégrade au cours du temps. Dietz *et al.* (2013); Kenter *et al.* (2015) ont proposé des nouvelles mesures d'évaluation capables d'identifier la dégradation de la performance d'un système au cours du temps. L'évaluation de nos propositions avec ces mesures devrait être menée.

- Ensuite, des travaux tels que le système *BIT_Purdue* (Wang *et al.*, 2014) exploitent des critères basés sur les verbes d'actions. Nous pensons adapter notre méthode basée sur les modèles de langue en estimant la vitalité à partir des verbes d'action ou d'autre termes pouvant mieux détecter la vitalité.

- Un troisième point concerne la stratégie de filtrage des documents vitaux. Dans nos méthodes, nous avons calculé un score de vitalité pour chaque document filtré (mentionnant l'entité et non rejeté par les filtres de spams). Nous pensons que la définition d'un seuil minimum pourrait améliorer la précision de notre système. Cependant, la valeur de ce seuil présente une problématique à résoudre. En effet, ce seuil peut varier selon l'équation du score de vitalité utilisée (l'équation 4.5 de la méthode 1, ou l'équation 4.9 de la méthode 2), ou encore en fonction de chaque entité.

- Un quatrième point concerne l'estimation de la nouveauté. En effet, nous décidons la nouveauté d'une phrase en la comparant à toutes les autres phrases déjà sélectionnées. Cette façon de faire ne considère que l'historique des phrases sélectionnées. Nous pensons que la détection des termes émergents (en rafale) dans le flux peut améliorer la détection de la nouveauté. Ces termes peuvent être détectés en exploitant certaines statistiques sur les termes dans le flux telles que le facteur *IDF* basé sur la fenêtre du temps précédant l'arrivée du document.

- Dans notre approche de génération de résumés temporels, nous nous sommes intéressés particulièrement à la pertinence et à la nouveauté des phrases. Notre perspective, à plus long terme, est de considérer d'autres dimensions telles que la complémentarité entre les phrases, la crédibilité (Weerkamp et de Rijke, 2008) et la lisibilité (Polajnar *et al.*, 2012).

Annexes

Annexe

Topics proposés dans le cadre de la tâche TREC KBA 2013 et 2014

TABLE 6.1: Topics proposés dans TREC KBA 2013. Nous utilisant l'abbréviation WP pour remplacer le début de l'URL *l'URL http ://en.wikipedia.org/wiki*, et TWT pour remplacer le début de l'URL *https ://twitter.com*

Topic	**Type**	**Group**
WP/Appleton_Museum_of_Art	FAC	ocala
WP/Copper_Basin_Railway	FAC	mining
WP/Corn_Belt_Power_Cooperative	FAC	fargo
WP/Don_Garlits_Museum_of_Drag_Racing	FAC	ocala
WP/Eighth_Street_Elementary_School	FAC	ocala
WP/Elysian_Charter_School	FAC	hoboken
WP/Fargo_Air_Museum	FAC	fargo
WP/Fargo-Moorhead_Symphony_Orchestra	FAC	fargo
WP/Hayden_Smelter	FAC	mining
WP/Hjemkomst_Center	FAC	fargo
WP/IDSIA	FAC	deeplearning
WP/John_D._Odegard_School_of_Aerospace_Sciences	FAC	fargo
WP/Lake_Weir_High_School	FAC	ocala
WP/Lewis_and_Clark_Landing	FAC	mining
WP/Marion_Technical_Institute	FAC	ocala
WP/Osceola_Middle_School	FAC	ocala
WP/Red_River_Zoo	FAC	fargo
WP/Star_Lite_Motel	FAC	fargo
WP/Stevens_Cooperative_School	FAC	hoboken
WP/Stuart_Powell_Field	FAC	danville
WP/The_Ritz_Apartment_(Ocala,_Florida)	FAC	ocala
WP/Toquepala_mine	FAC	mining
WP/Weehawken_Cove	FAC	hoboken
TWT/CorbinSpeedway	FAC	danville

WP/Agroindustrial_Pomalca	ORG	mining
WP/Atacocha	ORG	mining
WP/Austral_Group	ORG	mining
WP/Cementos_Lima	ORG	mining
WP/Dunkelvolk	ORG	mining
WP/Grana_y_Montero		
WP/Great_American_Brass_Band_Festival	ORG	danville
WP/Hoboken_Reporter	ORG	hoboken
WP/Hoboken_Volunteer_Ambulance_Corps	ORG	hoboken
WP/Innovis_Health	ORG	fargo
WP/Intergroup_Financial_Services	ORG	mining
WP/Luz_del_Sur	ORG	mining
WP/Scotiabank_Peru	ORG	mining
WP/SIMSA	ORG	mining
TWT/BlossomCoffee	ORG	startups
TWT/evvnt	ORG	startups
TWT/FrankandOak	ORG	startups
TWT/GandBcoffee	ORG	startups
TWT/WCoffeeResearch	ORG	startups
WP/Anaïs_Croze	PER	french
WP/Angelo_Savoldi	PER	hoboken
WP/Barbara_Liskov	PER	turing
WP/Benjamin_Bronfman	PER	bronfman
WP/Bernard_Kenny	PER	hoboken
WP/Blair_Thoreson	PER	fargo
WP/Bob_Bert	PER	hoboken
WP/Brenda_Weiler	PER	fargo
WP/Buddy_MacKay	PER	ocala
WP/Carey_McWilliams_(marksman)	PER	fargo
WP/Carl_Chang_(tennis)	PER	hoboken
WP/Carla_Katz	PER	hoboken
WP/Charles_Bronfman	PER	bronfman
WP/Chiara_Nappi	PER	hep
WP/Chuck_Pankow	PER	ocala
WP/Clare_Bronfman	PER	bronfman
WP/Clark_Blaise	PER	fargo
WP/Daniel_J._Crothers	PER	fargo
WP/Danny_Irmen	PER	fargo
WP/David_B._Danbom	PER	fargo
WP/DeAnne_Smith	PER	commedians
WP/Derrick_Alston	PER	hoboken
WP/Drew_Wrigley	PER	fargo
WP/Ed_Bok_Lee	PER	fargo

WP/Edgar_Bronfman,_Jr.	PER	bronfman
WP/Edgar_Bronfman,_Sr.	PER	bronfman
WP/Eva_Silverstein	PER	hep
WP/Fargo_Moorhead_Derby_Girls	PER	fargo
WP/Fernando_J._Corbató	PER	turing
WP/Frank_Winters	PER	hoboken
WP/Geoffrey_E._Hinton	PER	deeplearning
WP/George_Sinner	PER	fargo
WP/Gretchen_Hoffman	PER	fargo
WP/Gwenaëlle_Aubry	PER	french
WP/Haven_Denney	PER	ocala
WP/Henry_Gutierrez	PER	hoboken
WP/Jack_Lazorko	PER	hoboken
WP/Jamie_Parsley	PER	fargo
WP/Jasper_Schneider	PER	fargo
WP/Jeff_Severson	PER	fargo
WP/Jeff_Tamarkin	PER	hoboken
WP/Jennifer_Baumgardner	PER	fargo
WP/Jeremy_McKinnon	PER	ocala
WP/Jim_Poolman	PER	fargo
WP/Joanne_Borgella	PER	hoboken
WP/Joey_Mantia	PER	ocala
WP/John_H._Lang	PER	fargo
WP/Joshua_Boschee	PER	fargo
WP/Joshua_Zetumer	PER	screenwriters
WP/Judd_Davis	PER	ocala
WP/Juris_Hartmanis	PER	turing
WP/Ken_Fowler	PER	fargo
WP/Ken_Freedman	PER	hoboken
WP/Keri_Hehn	PER	fargo
WP/Klaus_Grutzka	PER	hoboken
WP/Léon_Bottou	PER	deeplearning
WP/Lorenzo_Williams_(basketball)	PER	ocala
WP/Mark_SaFranko	PER	hoboken
WP/Matt_Witten	PER	hep
WP/Maurice_Fitzgibbons	PER	hoboken
WP/Nicolas_Schöffer		
WP/Olaus_Murie	PER	fargo
WP/Pat_Dapuzzo	PER	hoboken
WP/Paul_Johnsgard	PER	fargo
WP/Paul_Marquart	PER	fargo
WP/Phyllis_Lambert	PER	bronfman
WP/Randy_Ewers	PER	ocala

WP/Reid_Nichols	PER	ocala
WP/Richard_Edlund	PER	fargo
WP/Richard_W._Goldberg	PER	fargo
WP/Ruben_J._Ramos	PER	hoboken
WP/Sara_Bronfman	PER	bronfman
WP/Scot_Brantley	PER	ocala
WP/Sean_Hampton	PER	ocala
WP/Shafi_Goldwasser	PER	turing
WP/Shamit_Kachru	PER	hep
WP/Susan_Krieg	PER	fargo
WP/Théo_Mercier	PER	french
WP/Tilo_Rivas	PER	hoboken
WP/Travis_Mays	PER	ocala
WP/William_H._Miller_(writer)	PER	hoboken
WP/William_P._Gerberding	PER	fargo
WP/Yann_LeCun	PER	deeplearning
WP/Zoubin_Ghahramani	PER	deeplearning
TWT/AlexJoHamilton	PER	danville
TWT/BartowMcDonald	PER	ocala
TWT/bobplaisted	PER	ocala
TWT/BobStovall	PER	danville
TWT/danvillekyengr	PER	danville
TWT/KentGuinn4Mayor	PER	ocala
TWT/MissMarcel	PER	screenwriters
TWT/redmondmusic	PER	ocala
TWT/RobCaud	PER	danville
TWT/RonFunches	PER	commedians
TWT/roryscovel	PER	commedians
TWT/sandrafriend	PER	ocala
TWT/tonyg203	PER	danville
TWT/urbren00	PER	danville

Table 6.2: Topics proposés dans TREC KBA 2014.

Topic	**Type**
Topic	Type
https ://kb.diffeo.com/Joby_Shimomura	PER
https ://kb.diffeo.com/Ted_Sturdevant	PER
https ://kb.diffeo.com/Dan_Satterberg	PER
https ://kb.diffeo.com/_3NIbfDpEdTwZ	PER
https ://kb.diffeo.com/BNSF_Railway	ORG
https ://kb.diffeo.com/IslandWood	FAC
https ://kb.diffeo.com/Jonathan_Meline	PER

https ://kb.diffeo.com/Dow_Constantine	PER
https ://kb.diffeo.com/Kshama_Sawant	PER
https ://kb.diffeo.com/Brodie_Clowes	PER
https ://kb.diffeo.com/Mark_Lindquist	PER
https ://kb.diffeo.com/Genaveve_Starr	PER
https ://kb.diffeo.com/Chad_Kroeger	PER
https ://kb.diffeo.com/Spokane_County_Raceway	FAC
https ://kb.diffeo.com/King_Cat_Theater	FAC
https ://kb.diffeo.com/Andy_Billig	PER
https ://kb.diffeo.com/James_Windle	PER
https ://kb.diffeo.com/Peter_Goldmark	PER
https ://kb.diffeo.com/Ralph_Dannenberg	PER
https ://kb.diffeo.com/J._Tillman	PER
https ://kb.diffeo.com/Georgie_Bright_Kunkel	PER
https ://kb.diffeo.com/Rick_Hansen	PER
https ://kb.diffeo.com/Carmela_Dellino	PER
https ://kb.diffeo.com/Paul_Watson	PER
https ://kb.diffeo.com/Semiahmoo_First_Nation	ORG
https ://kb.diffeo.com/Jean_Luc_Bilodeau	PER
https ://kb.diffeo.com/Tulalip	ORG
https ://kb.diffeo.com/Kalispel_Tribe	ORG
https ://kb.diffeo.com/Marty_McLaren	PER
https ://kb.diffeo.com/Damien_Jurado	PER
https ://kb.diffeo.com/Jacob_Hoggard	PER
https ://kb.diffeo.com/Spokane_Tribe	ORG
https ://kb.diffeo.com/Leona_Aglukkaq	PER
https ://kb.diffeo.com/Josh_Vander_Vies	PER
https ://kb.diffeo.com/Maelle_Ricker	PER
https ://kb.diffeo.com/Randy_Dorn	PER
https ://kb.diffeo.com/Spokane_Convention_Center	FAC
https ://kb.diffeo.com/Bryce_Leavitt	PER
https ://kb.diffeo.com/Mike_Kluse	PER
https ://kb.diffeo.com/Jim_Busey	PER
https ://kb.diffeo.com/Abbotsford_Arts_Centre	FAC
https ://kb.diffeo.com/Paul_Brandt	PER
https ://kb.diffeo.com/Jesse_Sykes	PER
https ://kb.diffeo.com/Nordic_Heritage_Museum	ORG
https ://kb.diffeo.com/Jeff_Mangum	PER
https ://kb.diffeo.com/Anne_Blair	PER
https ://kb.diffeo.com/Shelley_Redinger	PER
https ://kb.diffeo.com/matt_manweller	PER
https ://kb.diffeo.com/Mason_Wilgosh	PER
https ://kb.diffeo.com/Nolan_Watson	PER

https ://kb.diffeo.com/Snohomish_High_School	ORG
https ://kb.diffeo.com/Theresa_Spence	PER
https ://kb.diffeo.com/Tsleil-Waututh_First_Nation	ORG
https ://kb.diffeo.com/Brock_Schuh	PER
https ://kb.diffeo.com/Shawn_Atleo	PER
https ://kb.diffeo.com/Elmer_Derrick	PER
https ://kb.diffeo.com/Missing_Women_Commission_of_Inquiry	ORG
https ://kb.diffeo.com/Lisa_Brown	PER
https ://kb.diffeo.com/Tsawwassen_First_Nation	ORG
https ://kb.diffeo.com/Susan_Chapelle	PER
https ://kb.diffeo.com/Ted_Prior	PER
https ://kb.diffeo.com/Lizette_Graden	PER
https ://kb.diffeo.com/Bryan_Raiser	PER
https ://kb.diffeo.com/Nancy_Wilhelm_Morden	PER
https ://kb.diffeo.com/Lisa_Muri	PER
https ://kb.diffeo.com/Rob_Kirkham	PER
https ://kb.diffeo.com/Robyn_Gervais	PER
https ://kb.diffeo.com/Corisa_Bell	PER
https ://kb.diffeo.com/Cameron_Ward	PER
https ://kb.diffeo.com/_Bill_Templeton	PER
https ://kb.diffeo.com/Stephen_Buxbaum	PER
https ://kb.diffeo.com/Jessie_Kaech	PER
https ://kb.diffeo.com/Rose_Egge	PER
https ://kb.diffeo.com/_Dave_Rosin	PER
https ://kb.diffeo.com/A._J._Rathbun	PER
https ://kb.diffeo.com/Aaron_Yeung	PER
https ://kb.diffeo.com/Allison_Campbell	PER
https ://kb.diffeo.com/Andy_Clausen	PER
https ://kb.diffeo.com/Candace_Pratt	PER
https ://kb.diffeo.com/Captain_Wiggette	PER
https ://kb.diffeo.com/Chris_Boyd	PER
https ://kb.diffeo.com/Doug_Race	PER
https ://kb.diffeo.com/First_Nations_Summit	ORG
https ://kb.diffeo.com/Ian_Kent	PER
https ://kb.diffeo.com/Jamie_Lawson	PER
https ://kb.diffeo.com/Jud_Virden	PER
https ://kb.diffeo.com/Karen_Guzak	PER
https ://kb.diffeo.com/Karen_Pohl	PER
https ://kb.diffeo.com/Katzie_First_Nation	ORG
https ://kb.diffeo.com/Kendra_Obom	PER
https ://kb.diffeo.com/Kerry_Morris	PER
https ://kb.diffeo.com/Lake_Sammamish_Elks_Lodge	FAC
https ://kb.diffeo.com/Lisa_McCullough	PER

https ://kb.diffeo.com/Matheson_Farms	ORG
https ://kb.diffeo.com/Michael_K_Young	PER
https ://kb.diffeo.com/Muckleshoot_Indian_Tribe	ORG
https ://kb.diffeo.com/Nathaniel_Jones	PER
https ://kb.diffeo.com/Nez_Perse	ORG
https ://kb.diffeo.com/Patricia_Heintzman	PER
https ://kb.diffeo.com/Phil_Wandscher	PER
https ://kb.diffeo.com/Philippe_Cury	PER
https ://kb.diffeo.com/Rashid_Sumaila	PER
https ://kb.diffeo.com/Ron_Sander	PER
https ://kb.diffeo.com/Shelli_Park	PER
https ://kb.diffeo.com/Sun_Villa_Lanes	FAC
https ://kb.diffeo.com/Val_Tollefson	PER
https ://kb.diffeo.com/William_Azaroff	PER
https ://kb.diffeo.com/Women_Winemakers	ORG
https ://kb.diffeo.com/_Nicholas_Tse	PER

Bibliographie

Abbes, R., Hernandez, N., Pinel-Sauvagnat, K. et Boughanem, M. (2015a). Accelerating the update of knowledge base instances by detecting vital information from a document stream. *In IEEE/WIC/ACM International Conference on Web Intelligence, Singapour, 06/12/2015-09/12/2015*, http ://www.ieee.org/. IEEE.

Abbes, R., Hernandez, N., Pinel-Sauvagnat, K. et Boughanem, M. (2015b). Détection d'informations vitales pour la mise à jour de bases de connaissances. *In Journées Francophones d'Ingénierie des Connaissances (IC), Rennes, 29/06/15-03/07/15*, pages 147–158, http ://www.afia.asso.fr/. Association Francaise d'Intelligence Artificielle (AFIA).

Abbes, R., Pinel-Sauvagnat, K., Hernandez, N. et Boughanem, M. (2013). IRIT at TREC knowledge base acceleration 2013 : Cumulative citation recommendation task. *In Proceedings of The Twenty-Second Text REtrieval Conference, TREC 2013, Gaithersburg, Maryland, USA, November 19-22, 2013*.

Abbes, R., Pinel-Sauvagnat, K., Hernandez, N. et Boughanem, M. (2014a). IRIT at TREC KBA 2014. *In Proceedings of The Twenty-Third Text REtrieval Conference, TREC 2014, Gaithersburg, Maryland, USA, November 19-21, 2014*.

Abbes, R., Pinel-Sauvagnat, K., Hernandez, N. et Boughanem, M. (2014b). Modèles de langue pour la mise à jour d'un profil d'entité. *In CORIA 2014 - Conférence en Recherche d'Infomations et Applications-11th French Information Retrieval Conference. CIFED 2014 Colloque International Francophone sur l'Ecrit et le Document, Nancy, France, March 19-23, 2014.*, pages 137–151. ARIA-GRCE.

Abbes, R., Pinel-Sauvagnat, K., Hernandez, N. et Boughanem, M. (2015c). Leveraging temporal expressions to filter vital documents related to an entity. *In Proceedings of the 30th Annual ACM Symposium on Applied Computing*, SAC '15, pages 1093–1098, New York, NY, USA. ACM.

Abbes, R., Pinel-Sauvagnat, K., Hernandez, N. et Boughanem, M. (2015d). When temporal expressions help to detect vital documents related to an entity. *SIGAPP Appl. Comput. Rev.*, 15(3):49–58.

Allan, J., Carbonell, J., Doddington, G., Yamron, J. et Yang, Y. (1998a). Topic detection and tracking pilot study : Final report. *In Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*, pages 194–218, Lansdowne, VA, USA. 007.

Allan, J., Papka, R. et Lavrenko, V. (1998b). On-line new event detection and tracking. *In Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, pages 37–45, New York, NY, USA. ACM.

Allan, J., Wade, C. et Bolivar, A. (2003). Retrieval and novelty detection at the sentence level. *In Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, SIGIR '03, pages 314–321, New York, NY, USA. ACM.

Alonso, O., Baeza-yates, R., Strötgen, J. et Gertz, M. (2011). Temporal information retrieval : Challenges and opportunities. *In 1st Temporal Web Analytics Workshop at WWW*, pages 1–8.

Araújo, S., Gebremeskel, G. G., He, J., Boscarino, C. et de Vries, A. P. (2012). CWI at TREC 2012, KBA track and session track. *In Proceedings of The Twenty-First Text REtrieval Conference, TREC 2012, Gaithersburg, Maryland, USA, November 6-9, 2012.*

Aslam, J., Diaz, F., Ekstrand-Abueg, M., Pavlu, V. et Sakai, T. (2013). Trec 2013 temporal summarization. *In Proceedings of the Twenty-Second Text REtrieval Conference*, Gaithersburgh, USA.

Aslam, J. A., Ekstrand-Abueg, M., Pavlu, V., Diaz, F., McCreadie, R. et Sakai, T. (2014). TREC 2014 temporal summarization track overview. *In Proceedings of The Twenty-Third Text REtrieval Conference, TREC 2014, Gaithersburg, Maryland, USA, November 19-21, 2014.*

Augenstein, I., Padó, S. et Rudolph, S. (2012). Lodifier : Generating linked data from unstructured text. *In Proceedings of the 9th International Conference on The Semantic Web : Research and Applications*, ESWC'12, pages 210–224, Berlin, Heidelberg.

Baeza-Yates, R. A. et Ribeiro-Neto, B. A. (2011). *Modern Information Retrieval - the concepts and technology behind search, Second edition.* Pearson Education Ltd., Harlow, England.

Balog, K., de Vries, A. P., Serdyukov, P., Thomas, P. et Westerveld, T. (2009). Overview of the TREC 2009 entity track. *In Proceedings of The Eighteenth Text REtrieval Conference, TREC 2009, Gaithersburg, Maryland, USA, November 17-20, 2009*.

Balog, K. et Ramampiaro, H. (2013). Cumulative citation recommendation : Classification vs. ranking. *In Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '13, pages 941–944, New York, NY, USA. ACM.

Balog, K., Ramampiaro, H., Takhirov, N. et Nørvåg, K. (2013). Multi-step classification approaches to cumulative citation recommendation. *In Proceedings of the 10th Conference on Open Research Areas in Information Retrieval*, OAIR '13, pages 121–128, Paris, France.

Balog, K., Serdyukov, P. et de Vries, A. P. (2010). Overview of the TREC 2010 entity track. *In Proceedings of The Nineteenth Text REtrieval Conference, TREC 2010, Gaithersburg, Maryland, USA, November 16-19, 2010*.

Baruah, G., Guttikonda, R., Roegiest, A. et Vechtomova, O. (2013). University of waterloo at the TREC 2013 temporal summarization track. *In Proceedings of The Twenty-Second Text REtrieval Conference, TREC 2013, Gaithersburg, Maryland, USA, November 19-22, 2013*.

Belkin, N. J. et Croft, W. B. (1992). Information filtering and information retrieval : Two sides of the same coin ? *Commun. ACM*, 35(12):29–38.

Bizer, C., Heath, T. et Berners-Lee, T. (2009). Linked data - the story so far. *Int. J. Semantic Web Inf. Syst.*, 5(3):1–22.

Bollacker, K., Evans, C., Paritosh, P., Sturge, T. et Taylor, J. (2008). Freebase : A collaboratively created graph database for structuring human knowledge. *In Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, SIGMOD '08, pages 1247–1250, New York, NY, USA. ACM.

Bonnefoy, L., Bouvier, V. et Bellot, P. (2013). A weakly-supervised detection of entity central documents in a stream. *In Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '13, pages 769–772, New York, NY, USA. ACM.

Boughanem, M. (1992). *Les systèmes de recherche d'informations d'un modèle classique à un modèle connexioniste*. Thèse de doctorat, Universitè Paul Sabatier, Toulouse, France.

Boughanem, M., Kraaij, W. et Nie, J.-Y. (2004). Modèles de langue pour la recherche d'information. *In Les systèmes de recherche d'informations*, pages 163–182. Hermes-Lavoisier, Lavoisier, 11, rue Lavoisier 75008.

Bouvier, V. et Bellot, P. (2013). Filtering entity centric documents using numerics and temporals features within RF classifier. *In Proceedings of The Twenty-Second Text REtrieval Conference, TREC 2013, Gaithersburg, Maryland, USA, November 19-22, 2013*.

Bouvier, V. et Bellot, P. (2015). Regroupement par popularité pour la RI semi-supervisée centrée sur les entités. *In CORIA 2015 - Conférence en Recherche d'Infomations et Applications - 12th French Information Retrieval Conference, Paris, France, March 18-20, 2015.*, pages 503–512. ARIA.

Campos, R., Dias, G., Jorge, A. M. et Jatowt, A. (2014). Survey of temporal information retrieval and related applications. *ACM Comput. Surv.*, 47(2):15 :1–15 :41.

Cao, G., Nie, J.-Y., Gao, J. et Robertson, S. (2008). Selecting good expansion terms for pseudo-relevance feedback. *In Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '08, pages 243–250, New York, NY, USA. ACM.

Carenini, G. et Cheung, J. C. K. (2008). Extractive vs. nlg-based abstractive summarization of evaluative text : The effect of corpus controversiality. *In Proceedings of the Fifth International Natural Language Generation Conference*, INLG '08, pages 33–41, Stroudsburg, PA, USA. Association for Computational Linguistics.

Chang, A. X. et Manning, C. (2012). Sutime : A library for recognizing and normalizing time expressions. *In Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Chen, L., Zhang, H., Li, S., Ji, Z., Liu, Q., Liu, Y., Wu, D. et Cheng, X. (2014). ICTNET at temporal summarization track TREC 2014. *In Proceedings of The Twenty-Third Text REtrieval Conference, TREC 2014, Gaithersburg, Maryland, USA, November 19-21, 2014*.

Chen, P. P.-S. (1976). The entity-relationship model—toward a unified view of data. *ACM Trans. Database Syst.*, 1(1):9–36.

Chowdhury, G. (2010). *Introduction to Modern Information Retrieval, Third Edition*. Facet Publishing, 3rd édition.

Craswell, N., Hawking, D., Vercoustre, A.-M. et Wilkins, P. (2001). Pnoptic expert : Searching for experts not just for documents. *In In Ausweb*, pages 21–25.

Cucerzan, S. (2007). Large-scale named entity disambiguation based on wikipedia data. *In EMNLP-CoNLL 2007, Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, June 28-30, 2007, Prague, Czech Republic*, pages 708–716.

Dang, H. T. et Owczarzak, K. (2008). Overview of the TAC 2008 update summarization task. *In Proceedings of the First Text Analysis Conference, TAC 2008, Gaithersburg, Maryland, USA, November 17-19, 2008*.

Dietz, L. et Dalton, J. (2013). Umass at TREC KBA 2013. TREC'13, Gaithersburg, USA.

Dietz, L., Dalton, J. et Balog, K. (2013). Time-aware evaluation of cumulative citation recommendation systems. *In SIGIR 2013 Workshop on Time-aware Information Access (TAIA2013)*.

Dueck, D. et Frey, B. J. (2007). Non-metric affinity propagation for unsupervised image categorization. *In ICCV*, pages 1–8. IEEE.

Efron, M., Willis, C. et Sherman, G. (2014). Learning sufficient queries for entity filtering. *In Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '14, pages 1091–1094, New York, NY, USA. ACM.

Erkan, G. et Radev, D. R. (2004). Lexrank : Graph-based lexical centrality as salience in text summarization. *J. Artif. Int. Res.*, 22(1):457–479.

Exner, P. et Nugues, P. (2012). Entity extraction : From unstructured text to dbpedia rdf triples. WoLE'12.

Filatova, E. et Hatzivassiloglou, V. (2004). A formal model for information selection in multi-sentence text extraction. *In Proceedings of the 20th International Conference on Computational Linguistics*, COLING '04, Stroudsburg, PA, USA. Association for Computational Linguistics.

Frank, J. R., Bauer, S. J., Kleiman-Weiner, M., Roberts, D. A., Tripuraneni, N., Zhang, C., Ré, C., Voorhees, E. M. et Soboroff, I. (2013). Evaluating stream filtering for entity profile updates for TREC 2013. *In Proceedings of The Twenty-Second Text REtrieval Conference, TREC 2013, Gaithersburg, Maryland, USA, November 19-22, 2013*. National Institute of Standards and Technology (NIST).

Frank, J. R., Kleiman-Weiner, M., Roberts, D. A., Niu, F., Zhang, C., Re, C. et Soboroff, I. (2012). Building an Entity-Centric stream filtering test collection for TREC 2012. *In Proceedings of the Text REtrieval Conference (TREC)*.

Frank, J. R., Kleiman-Weiner, M., Roberts, D. A., Voorhees, E. M. et Soboroff, I. (2014). Evaluating stream filtering for entity profile updates in TREC 2012, 2013, and 2014. *In Proceedings of The Twenty-Third Text REtrieval Conference, TREC 2014, Gaithersburg, Maryland, USA, November 19-21, 2014*. National Institute of Standards and Technology (NIST).

Franz, M., Ittycheriah, A., McCarley, J. S. et Ward, T. (2001). First story detection : Combining similarity and novelty-based approaches.

Furnas, G. W., Deerwester, S., Dumais, S. T., Landauer, T. K., Harshman, R. A., Streeter, L. A. et Lochbaum, K. E. (1988). Information retrieval using a singular value decomposition model of latent semantic structure. *In Proceedings of the 11th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '88, pages 465–480, New York, NY, USA. ACM.

Ganesan, K., Zhai, C. et Han, J. (2010). Opinosis : A graph based approach to abstractive summarization of highly redundant opinions. *In Proceedings of the 23rd International Conference on Computational Linguistics (COLING '10)*.

Goldberg, D., Nichols, D., Oki, B. M. et Terry, D. (1992). Using collaborative filtering to weave an information tapestry. *Commun. ACM*, 35(12):61–70.

Guo, J., Xu, G., Cheng, X. et Li, H. (2009). Named entity recognition in query. *In Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, pages 267–274, New York, NY, USA. ACM.

Harman, D. (2002). Overview of the TREC 2002 novelty track. *In Proceedings of The Eleventh Text REtrieval Conference, TREC 2002, Gaithersburg, Maryland, USA, November 19-22, 2002*.

He, Q., Chang, K., Lim, E.-P. et 0005, J. Z. (2007). Bursty feature representation for clustering text streams. *In SDM*, pages 491–496.

Herlocker, J. L., Konstan, J. A., Borchers, A. et Riedl, J. (1999). An algorithmic framework for performing collaborative filtering. *In Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '99, pages 230–237, New York, NY, USA. ACM.

Hirschman, L. et Gaizauskas, R. (2001). Natural language question answering : The view from here. *Nat. Lang. Eng.*, 7(4):275–300.

Hovy, E. et Lin, C.-Y. (1998a). Automated text summarization and the summarist system. *In Proceedings of a Workshop on Held at Baltimore, Maryland : October 13-15, 1998*, TIPSTER '98, pages 197–214, Stroudsburg, PA, USA. Association for Computational Linguistics.

Hovy, E. et Lin, C.-Y. (1998b). Automated text summarization and the summarist system. *In Proceedings of a Workshop on Held at Baltimore, Maryland : October 13-15, 1998*, TIPSTER '98, pages 197–214, Stroudsburg, PA, USA. Association for Computational Linguistics.

Hull, D. A. (1998). The TREC-7 filtering track : Description and analysis. *In Proceedings of The Seventh Text REtrieval Conference, TREC 1998, Gaithersburg, Maryland, USA, November 9-11, 1998*, pages 9–32.

Hull, D. A. et Robertson, S. E. (1999). The TREC-8 filtering track final report. *In Proceedings of The Eighth Text REtrieval Conference, TREC 1999, Gaithersburg, Maryland, USA, November 17-19, 1999.*

Järvelin, K. et Kekäläinen, J. (2002). Cumulated gain-based evaluation of ir techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446.

Jelinek, F. et Mercer, R. L. (1980). Interpolated estimation of Markov source parameters from sparse data. *In* Gelsema, E. S. et Kanal, L. N., éditeurs : *Proceedings, Workshop on Pattern Recognition in Practice*, pages 381–397. North Holland, Amsterdam.

Jiang, J., Lin, C.-Y. et Rui, Y. (2014). Msr kmg at trec 2014 kba track vital filtering task. *In Notebook of The Twenty-Third Text REtrieval Conference, TREC 2014, Gaithersburg, Maryland, USA, November 19-21, 2014.*

Jones, K. S. et van Rijsbergen, C. J. (1975). Report on the Need for and Provision of an "Ideal" Information Retrieval Test Collection. British Library Research and Development Report 5266, University of Cambridge.

Karkali, M., Rousseau, F., Ntoulas, A. et Vazirgiannis, M. (2014). Using temporal IDF for efficient novelty detection in text streams. *CoRR*, abs/1401.1456.

Katz, S. (1987). Estimation of probabilities from sparse data for the language model component of a speech recognizer. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 35(3):400–401.

Kazawa, H., Hirao, T., Isozaki, H. et Maeda, E. (2002). A machine learning approach for QA and novelty tracks : NTT system description.

In Proceedings of The Eleventh Text REtrieval Conference, TREC 2002, Gaithersburg, Maryland, USA, November 19-22, 2002.

Kedzie, C., McKeown, K. et Diaz, F. (2014). Columbia university at TREC 2014 : Temporal summarization. *In Proceedings of The Twenty-Third Text REtrieval Conference, TREC 2014, Gaithersburg, Maryland, USA, November 19-21, 2014*.

Kenter, T. (2013). Filtering documents over time on evolving topics - the university of amsterdam at TREC 2013 KBA CCR. *In Proceedings of The Twenty-Second Text REtrieval Conference, TREC 2013, Gaithersburg, Maryland, USA, November 19-22, 2013*.

Kenter, T., Balog, K. et de Rijke, M. (2015). Evaluating document filtering systems over time. *Information Processing & Management*, pages –.

Kleinberg, J. (2002). Bursty and hierarchical structure in streams. *In Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '02, pages 91–101, New York, NY, USA. ACM.

Kopliku, A., Boughanem, M. et Pinel-Sauvagnat, K. (2010). Querying by examples (poster). *In Conférence francophone en Recherche d'Information et Applications (CORIA), Sousse, Tunisie, 18/03/2010-20/03/2010*, pages 407–408, http ://www.irit.fr/ARIA. Association Francophone de Recherche d'Information et Applications (ARIA).

Kopliku, A., Pinel-Sauvagnat, K. et Boughanem, M. (2014). Aggregated search : A new information retrieval paradigm. *ACM Comput. Surv.*, 46(3):41 :1–41 :31.

Kullback, S. et Leibler, R. A. (1951). On information and sufficiency. *Ann. Math. Statist.*, 22(1):79–86.

Lang, K. (1995). Newsweeder : Learning to filter netnews. *In in Proceedings of the 12th International Machine Learning Conference (ML95*.

Larkey, L., Allan, J., Connell, M., Bolivar, A. et Wade, C. (2002). Umass at trec 2002 : Cross language and novelty tracks,. *In Proceedings (notebook version) of the TREC 2002*, pages 43–55.

Lavrenko, V. et Croft, W. B. (2001). Relevance based language models. *In Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '01, pages 120–127, New York, NY, USA. ACM.

Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P. N., Hellmann, S., Morsey, M., van Kleef, P., Auer, S. et Bizer, C. (2015). Dbpedia - A large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web*, 6(2):167–195.

Lin, C.-Y. et Hovy, E. (2002). From single to multi-document summarization : A prototype system and its evaluation. *In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 457–464, Stroudsburg, PA, USA. Association for Computational Linguistics.

Liu, Q., Liu, Y., Wu, D. et Xueqi, C. (2013a). Ictnet at temporal summarization track trec 2013. *In Proceedings of the Twenty-Second Text REtrieval Conference*, Gaithersburgh, MD, USA.

Liu, X., Fang, H. et Darko, J. (2013b). A related entity based approach for knowledge base acceleration. TREC'13, Gaithersburg, USA.

Long, R., Wang, H., Chen, Y., Jin, O. et Yu, Y. (2011). Towards effective event detection, tracking and summarization on microblog data. *In Proceedings of the 12th International Conference on Web-age Information Management*, WAIM'11, pages 652–663, Berlin, Heidelberg. Springer-Verlag.

Maña López, M. J., De Buenaga, M. et Gómez-Hidalgo, J. M. (2004). Multidocument summarization : An added value to clustering in interactive retrieval. *ACM Trans. Inf. Syst.*, 22(2):215–241.

Macdonald, C. et Ounis, I. (2006). Voting for candidates : Adapting data fusion techniques for an expert search task. *In Proceedings of the 15th ACM International Conference on Information and Knowledge Management*, CIKM '06, pages 387–396, New York, NY, USA. ACM.

MacKay, D. J. et Peto, L. C. B. (1994). A hierarchical dirichlet language model. *Natural Language Engineering*, 1:1–19.

Mackie, S., McCreadie, R., Macdonald, C. et Ounis, I. (2014). Comparing algorithms for microblog summarisation. *In Information Access Evaluation. Multilinguality, Multimodality, and Interaction - 5th International Conference of the CLEF Initiative, CLEF 2014, Sheffield, UK, September 15-18, 2014. Proceedings*, pages 153–159.

Malone, T. W., Grant, K. R., Turbak, F. A., Brobst, S. A. et Cohen, M. D. (1987). Intelligent information-sharing systems. *Commun. ACM*, 30(5):390–402.

Manning, C. D., Raghavan, P. et Schtze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA.

Maron, M. E. et Kuhns, J. L. (1960). On relevance, probabilistic indexing and information retrieval. *J. ACM*, 7(3):216–244.

Mayfield, J. et McNamee, P. (2003). Single n-gram stemming. *In Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, SIGIR '03, pages 415–416, New York, NY, USA. ACM.

McCreadie, R., Deveaud, R., Albakour, M., Mackie, S., Limsopatham, N., Macdonald, C., Ounis, I., Thonet, T. et Dinçer, B. T. (2014a). University of glasgow at TREC 2014 : Experiments with terrier in contextual suggestion, temporal summarisation and web tracks. *In Proceedings of The Twenty-Third Text REtrieval Conference, TREC 2014, Gaithersburg, Maryland, USA, November 19-21, 2014*.

McCreadie, R., Macdonald, C. et Ounis, I. (2014b). Incremental update summarization : Adaptive sentence selection based on prevalence and novelty. *In Proceedings of the 23rd International Conference on Conference on Information and Knowledge Management*, pages 301–310, New York, USA.

McKeown, K., Passonneau, R. J., Elson, D. K., Nenkova, A. et Hirschberg, J. (2005). Do summaries help ? *In Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '05, pages 210–217, New York, NY, USA. ACM.

Meij, E., Balog, K. et Odijk, D. (2014). Entity linking and retrieval for semantic search. *In Proceedings of the 7th ACM International Conference on Web Search and Data Mining*, WSDM '14, pages 683–684, New York, NY, USA. ACM.

Moulahi, B., Tamine, L. et Ben Yahia, S. (2015). When Time Meets Information Retrieval, Past Proposals, Current Plans, and Future Trends. *Journal of Information Science*.

Nenkova, A., Maskey, S. et Liu, Y. (2011). Automatic summarization. *In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Tutorial Abstracts of ACL 2011*, HLT '11, pages 3 :1–3 :86, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ney, H., Essen, U. et Kneser, R. (1995). On the estimation of 'small' probabilities by leaving-one-out. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 17(12):1202–1212.

Nie, Z., Wen, J. et Ma, W. (2012). Statistical entity extraction from the web. *Proceedings of the IEEE*, 100(9):2675–2687.

Ogawa, Y., Morita, T. et Kobayashi, K. (1991). A fuzzy document retrieval system using the keyword connection matrix and a learning method. *Fuzzy Sets Syst.*, 39(2):163–179.

Olariu, A. (2013). Hierarchical clustering in improving microblog stream summarization. *In Proceedings of the 14th International Conference on Computational Linguistics and Intelligent Text Processing - Volume 2*, CICLing'13, pages 424–435, Berlin, Heidelberg. Springer-Verlag.

Petasis, G., Karkaletsis, V., Paliouras, G., Krithara, A. et Zavitsanos, E. (2011). Ontology population and enrichment : State of the art. *In Knowledge-driven multimedia information extraction and ontology evolution*, pages 134–166. Springer-Verlag.

Polajnar, T., Glassey, R. et Azzopardi, L. (2012). Detection of news feeds items appropriate for children. *In Proceedings of the 34th European Conference on Advances in Information Retrieval*, ECIR'12, pages 63–72, Berlin, Heidelberg. Springer-Verlag.

Ponte, J. M. et Croft, W. B. (1998). A language modeling approach to information retrieval. *In Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, pages 275–281, New York, NY, USA. ACM.

Porter, M. F. (1997). Readings in information retrieval. chapitre An Algorithm for Suffix Stripping, pages 313–316. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Pound, J., Mika, P. et Zaragoza, H. (2010). Ad-hoc object retrieval in the web of data. *In Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 771–780, New York, NY, USA. ACM.

Radev, D. R., Jing, H., Styś, M. et Tam, D. (2004). Centroid-based summarization of multiple documents. *Information Processing and Management*, 40(6):919–938.

Robertson, S. E. et Hull, D. A. (2000). The TREC-9 filtering track final report. *In Proceedings of The Ninth Text REtrieval Conference, TREC 2000, Gaithersburg, Maryland, USA, November 13-16, 2000*.

Robertson, S. E. et Walker, S. (1999). Okapi/keenbow at TREC-8. *In Proceedings of The Eighth Text REtrieval Conference, TREC 1999, Gaithersburg, Maryland, USA, November 17-19, 1999*.

Robertson, S. E., Walker, S., Jones, S., Hancock-Beaulieu, M. et Gatford, M. (1994). Okapi at TREC-3. *In Proceedings of The Third Text REtrieval Conference, TREC 1994, Gaithersburg, Maryland, USA, November 2-4, 1994*, pages 109–126.

Salton, G. (1968). A comparison between manual and automatic indexing methods. Rapport technique, Ithaca, NY, USA.

Salton, G., éditeur (1971). *Relevance Feedback in Information Retrieval*. Prentice Hall, Englewood, Cliffs, New Jersey.

Salton, G. et McGill, M. J. (1986). *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., New York, NY, USA.

Salton, G., Wong, A. et Yang, C. S. (1975). A vector space model for automatic indexing. *Commun. ACM*, 18(11):613–620.

Sanderson, M. (2010). Test collection based evaluation of information retrieval systems. *Foundations and Trends® in Information Retrieval*, 4(4):247–375.

Saracevic, T. (1995). Evaluation of evaluation in information retrieval. *In Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '95, pages 138–146, New York, NY, USA. ACM.

Sharifi, B., Hutton, M.-A. et Kalita, J. (2010). Summarizing microblogs automatically. *In Human Language Technologies : The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 685–688, Stroudsburg, PA, USA. Association for Computational Linguistics.

Sherman, G., Efron, M. et Willis, C. (2014). The university of illinois' graduate school of library and information science at TREC 2014. *In Proceedings of The Twenty-Third Text REtrieval Conference, TREC 2014, Gaithersburg, Maryland, USA, November 19-21, 2014*. National Institute of Standards and Technology (NIST).

Singhal, A. (2012). Introducing the knowledge graph : things, not strings. *Official Google Blog, May*.

Soboroff, I. (2004). Overview of the TREC 2004 novelty track. *In Proceedings of the Thirteenth Text REtrieval Conference, TREC 2004, Gaithersburg, Maryland, USA, November 16-19, 2004*.

Soboroff, I. et Harman, D. (2003). Overview of the TREC 2003 novelty track. *In Proceedings of The Twelfth Text REtrieval Conference, TREC 2003, Gaithersburg, Maryland, USA, November 18-21, 2003*, pages 38–53.

Soboroff, I. et Harman, D. (2005). Novelty detection : The trec experience. *In Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 105–112, Stroudsburg, PA, USA. Association for Computational Linguistics.

Spitters, M. et Kraaij, W. (2001). TNO at TDT2001 : Language model-based topic detection. *In Topic Detection and Tracking (TDT) Workshop 2001*. NIST.

Surdeanu, M. et Heng, J. (2014). Overview of the english slot filling track at the tac2014 knowledge base population evaluation. *In Proceedings of the TAC-KBP 2014 Workshop*.

Tebri, H. (2004). *Formalisation et spécification d'un système de filtrage incrémental d'information*. Thèse de doctorat, Universitè Paul Sabatier, Toulouse, France.

Tmar, M. (2002). *Modele auto-adaptatif de filtrage d'information : apprentissage incremental du profil et de la fonction de decision*. Thèse de doctorat, Universitè Paul Sabatier, Toulouse, France.

Turtle, H. et Croft, W. B. (1990). Inference networks for document retrieval. *In Proceedings of the 13th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '90, pages 1–24, New York, NY, USA. ACM.

Vlachos, M., Meek, C., Vagena, Z. et Gunopulos, D. (2004). Identifying similarities, periodicities and bursts for online search queries. *In Proceedings of the 2004 ACM SIGMOD International Conference on Management of Data*, SIGMOD '04, pages 131–142, New York, NY, USA. ACM.

Voorhees, E. M. (2004). Overview of the TREC 2004 question answering track. *In Proceedings of the Thirteenth Text REtrieval Conference, TREC 2004, Gaithersburg, Maryland, USA, November 16-19, 2004*.

Vydiswaran, V. G. V., Ganesan, K., Lv, Y., He, J. et Zhai, C. (2009). Finding related entities by retrieving relations : UIUC at TREC 2009 entity track. *In Proceedings of The Eighteenth Text REtrieval Conference, TREC 2009, Gaithersburg, Maryland, USA, November 17-20, 2009*.

Wan, X. et Yang, J. (2008). Multi-document summarization using cluster-based link analysis. *In Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '08, pages 299–306, New York, NY, USA. ACM.

Wang, D. et Li, T. (2010). Document update summarization using incremental hierarchical clustering. *In Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, CIKM '10, pages 279–288, New York, NY, USA. ACM.

Wang, D., Li, T., Zhu, S. et Ding, C. (2008). Multi-document summarization via sentence-level semantic analysis and symmetric matrix factorization. *In Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '08, pages 307–314, New York, NY, USA. ACM.

Wang, J., Liao, L., Song, D., Ma, L., Lin, C. et Rui, Y. (2015a). Resorting relevance evidences to cumulative citation recommendation for knowledge base acceleration. *In Web-Age Information Management - 16th International Conference, WAIM 2015, Qingdao, China, June 8-10, 2015. Proceedings*, pages 169–180.

Wang, J., Song, D., Lin, C.-Y. et Liao, L. (2013). BIT and MSRA at TREC KBA CCR Track 2013. *In Proceedings of The Twenty-Second Text REtrieval Conference, TREC 2013, Gaithersburg, Maryland, USA, November 19-22, 2013*, TREC'13, Gaithersburg, USA.

Wang, J., Song, D., Wang, Q., Zhang, Z., Si, L., Liao, L. et Lin, C.-Y. (2015b). An entity class-dependent discriminative mixture model for cumulative citation recommendation. *In Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '15, pages 635–644, New York, NY, USA. ACM.

Wang, J., Zhang, N., Zhang, Z., Song, D., Si, L. et Liao, L. (2014). BIT and purdue at TREC-KBA-CCR track 2014. *In Proceedings of The Twenty-Third Text REtrieval Conference, TREC 2014, Gaithersburg, Maryland, USA, November 19-21, 2014*.

Weerkamp, W. et de Rijke, M. (2008). Credibility improves topical blog post retrieval. *In ACL*, pages 923–931.

Wong, S. K. M., Ziarko, W. et Wong, P. C. N. (1985). Generalized vector spaces model in information retrieval. *In Proceedings of the 8th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '85, pages 18–25, New York, NY, USA. ACM.

Xi, Y., Li, B., Zhou, J. et Tang, Y. (2013). ZZISTI at TREC2013 temporal summarization track. *In Proceedings of The Twenty-Second Text REtrieval Conference, TREC 2013, Gaithersburg, Maryland, USA, November 19-22, 2013*.

Xu, T., McNamee, P. et W.Qard, D. (2013a). HLTCOE at TREC 2013 : Temporal summarization. *In Proceedings of the Text REtrieval Conference*, Gaithersburgh.

Xu, T., Oard, D. W. et McNamee, P. (2013b). HLTCOE at TREC 2013 : Temporal summarization. *In Proceedings of The Twenty-Second Text REtrieval Conference, TREC 2013, Gaithersburg, Maryland, USA, November 19-22, 2013*.

Xu, Y., Jones, G. J. et Wang, B. (2009). Query dependent pseudo-relevance feedback based on wikipedia. *In Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, pages 59–66, New York, NY, USA. ACM.

Yang, Z., Yao, F., Sun, H., Zhao, Y., Lai, Y. et Fan, K. (2013). BJUT at TREC 2013 temporal summarization track. *In Proceedings of The Twenty-Second Text REtrieval Conference, TREC 2013, Gaithersburg, Maryland, USA, November 19-22, 2013*.

Zablith, F., Antoniou, G., d'Aquin, M., Flouris, G., Kondylakis, H., Motta, E., Plexousakis, D. et Sabou, M. (2015). Ontology evolution : a process-centric survey. *The Knowledge Engineering Review*, 30(01):45–75.

Zhai, C. et Lafferty, J. (2004). A study of smoothing methods for language models applied to information retrieval. *ACM Trans. Inf. Syst.*, 22(2):179–214.

Zhang, C., Xu, W., Meng, F., Li, H., Wu, T. et Xu, L. (2013). The information extracion systems of pris at temporal summarization track. *In Proceedings of the Text REtrieval Conference*, Gaithersburgh.

Zhang, H., Xu, H., Bai, S., Wang, B. et Cheng, X. (2004). Experiments in TREC 2004 novelty track at CAS-ICT. *In Proceedings of the Thirteenth Text REtrieval Conference, TREC 2004, Gaithersburg, Maryland, USA, November 16-19, 2004*.

Zhang, M., Song, R., Lin, C., Ma, S., Jiang, Z., Jin, Y., Liu, Y. et Zhao, L. (2002a). Expansion-based technologies in finding relevant and new information : Thu trec2002 novelty track experiments. *In Proceedings of The Eleventh Text REtrieval Conference, TREC 2002, Gaithersburg, Maryland, USA, November 19-22, 2002*.

Zhang, T., Ramakrishnan, R. et Livny, M. (1996). Birch : An efficient data clustering method for very large databases. *In Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data*, SIGMOD '96, pages 103–114, New York, NY, USA. ACM.

Zhang, Y., Callan, J. et Minka, T. (2002b). Novelty and redundancy detection in adaptive filtering. *In Proceedings of the 25th annual*

international ACM SIGIR conference on Research and development in information retrieval, SIGIR '02, pages 81–88, New York, NY, USA. ACM.

Zhao, Y., Yao, F., Sun, H. et Yang, Z. (2014). BJUT at TREC 2014 temporal summarization track. *In Proceedings of The Twenty-Third Text REtrieval Conference, TREC 2014, Gaithersburg, Maryland, USA, November 19-21, 2014*.

Zhou, M. et Chang, K. C.-C. (2013). Entity-centric document filtering : Boosting feature mapping through meta-features. *In Proceedings of the 22Nd ACM International Conference on Conference on Information & Knowledge Management*, CIKM '13, pages 119–128, New York, NY, USA. ACM.

Zhu, Y. et Shasha, D. (2003). Efficient elastic burst detection in data streams. *In Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '03, pages 336–345, New York, NY, USA. ACM.