De la recherche de granules documentaires à l'agrégation d'information

Karen Pinel-Sauvagnat

MANUSCRIT

En vue de l'obtention de l'

HABILITATION À DIRIGER DES RECHERCHES DE L'UNIVERSITÉ DE TOULOUSE

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

Présentée et soutenue le *13/07/2018* **par :**

Karen PINEL-SAUVAGNAT

De la recherche de granules documentaires à l'agrégation d'information

---

JURY

Nathalie Aussenac-Gilles	Directrice de Recherche CNRS, Univ. Toulouse 3	Examinatrice
Patrice Bellot	Professeur, Aix-Marseille Université	Rapporteur
Catherine Berrut	Professeure, Université Grenoble-Alpes	Rapporteuse
Mohand Boughanem	Professeur, Université Toulouse 3	Garant
Patrick Gallinari	Professeur, UPMC Paris	Rapporteur
Eric Gaussier	Professeur, Université Grenoble-Alpes	Examinateur

---

École doctorale et spécialité :

MITT : Image, Information, Hypermédia

Unité de Recherche :

IRIT - Institut de Recherche en Informatique de Toulouse - UMR 5505

- Et ta maman elle fait quoi comme métier ?
- Maman, elle est prof.
- Elle est professeure ?
- Non. Prof.
- D'accord. Prof de quoi ?
- Prof de danse.

Ellis P., 5 ans,
sur le métier d'enseignant-chercheur en informatique

"And what is the use of a book," thought Alice, "without pictures or conversation?"

Lewis Carroll, *Alice in Wonderland*

[Sauf mention contraire, les citations dans la suite de ce mémoire sont toutes extraites des œuvres de Lewis Carroll, *Alice in Wonderland* et *Through the looking glass*. Pourquoi ce choix ? Pour le non-sens et la logique qui leur sont propres, notre monde universitaire n'en étant parfois pas si loin.]

Remerciements

Ce mémoire présente l'essentiel des travaux de recherche que j'ai effectués au sein de l'IRIT depuis 2006. Il ne présente pas (ou très peu) de travaux individuels, et il n'existerait simplement pas sans toutes les personnes qui ont participé et contribué à ces recherches. C'est pourquoi je tiens à remercier en premier lieu tous les étudiants que j'ai eu la chance de co-encadrer, ainsi que Mohand Boughanem, qui m'a proposé ces co-encadrements, et qui, je dois bien le dire, a poussé pour l'écriture de ces pages.

Lobna, Mouna, Arlind, Cyril, Firas, Rafik, Thomas, Thibaut, MERCI. Autant de façons différentes d'encadrer et de travailler. J'espère vous avoir appris un peu, j'ai en tout cas appris beaucoup. Merci pour ces discussions constructives, vos qualités humaines, et ces années de travail commun. C'était un plaisir de travailler avec vous. Moh, MERCI. Merci de m'avoir emmenée jusque là, merci pour votre bonne humeur, toutes ces collaborations, et pour tout ce que j'ai appris à vos côtés sur la recherche et la direction de recherches.

Je remercie très sincèrement Catherine Berrut, professeure à l'Université Grenoble-Alpes, Patrice Bellot, professeur à l'université d'Aix-Marseille, et Patrick Gallinari, professeur à l'Université Pierre et Marie Curie Paris d'avoir accepté de rapporter sur ces travaux. Mes remerciements vont également aux examinateurs de ce jury, Eric Gaussier professeur à l'Université Grenoble-Alpes et Nathalie Aussenac-Gilles, directrice de recherche CNRS à l'Université de Toulouse. Merci à tous pour votre expertise et pour l'honneur que vous me faites de participer à mon jury.

Je tiens aussi et évidemment à remercier mes collègues de l'équipe IRIS, pour la bonne humeur quotidienne, pour toutes les collaborations passées et dans les tuyaux, et pour les discussions enflammées sur notre milieu universitaire souvent ubuesque. Les citations à la suite de ces remerciements sont pour vous. Je n'oublie pas, même s'ils ne sont pas cité nommément et même si ces remerciements sont maladroits, tous les collègues qui m'ont apporté leur aide et leur collaboration pour l'enseignement ou la recherche, ou qui ont tout simplement rendu le travail plus agréable.

Deux remerciements particuliers maintenant, à des collègues et amies outre-IRIS. Merci à Cécile pour son amitié et nos discussions sans fin de mamans de 3 enfants. Un merci immense à Nathalie, pour les travaux partagés bien sûr, mais surtout d'être là, encore et toujours.

Enfin, une dernière place spéciale pour mon mari et mes 3 gars, qui, même s'ils savent maintenant que je ne danse pas au travail, me font danser dans la vie.

“But I don’t want to go among mad people," Alice remarked.
Oh, you can’t help that," said the Cat: "we’re all mad here. I’m mad. You’re mad."
How do you know I’m mad?" said Alice.
You must be," said the Cat, "or you wouldn’t have come here.”

"Take some more tea," the March Hare said to Alice, very earnestly.
"I’ve had nothing yet," Alice replied in an offended tone, "so I can’t take more."
"You mean you can’t take less," said the Hatter: "it’s very easy to take more than nothing."

Table des matières

Liste des figures

Liste des tableaux

Introduction

> "Begin at the beginning," the King said, very gravely, "and go on till you come to the end: then stop."

Sommaire

Ce mémoire présente mes travaux de recherche dans le domaine de la *Recherche d'Information* (RI) depuis mon recrutement en tant que maître de conférences à l'Université Paul Sabatier (Toulouse 3) en 2006. Ces travaux ont été menés au sein de l'équipe SIG (*Systèmes d'Information Généralisés*) puis IRIS (*Information Retrieval and Information Synthesis*) de l'IRIT (*Institut de Recherche en Informatique de Toulouse*). L'équipe IRIS travaille autour des données complexes, hétérogènes et évolutives, selon deux directions : la recherche et la synthèse d'information.

Si l'organisation et la présentation de ce mémoire me sont propres, le contenu même des travaux décrits est le résultat d'un travail collaboratif et d'encadrement de thèses, qui n'aurait donc pas lieu d'être sans toutes les personnes avec qui j'ai travaillé et qui sont citées dans ces travaux.

Contexte des travaux : vers la recherche d'information agrégée

De cette définition de 1968 :

> An information retrieval system does not inform (i.e., change the knowledge of) the user on the subject of his inquiry. It merely informs on the existence (or non-existence) and whereabouts of documents relating to his request.
>
> F.W. Lancaster,
> *Information Retrieval Systems : Characteristics, Testing and Evaluation*

il ne reste aujourd'hui plus grand chose. L'expansion du Web depuis la fin des années 90 a modifié en profondeur le fonctionnement des *Systèmes de Recherche d'Information* (SRI).

Côté **documents**, des données sons, images et vidéos sont venues s'ajouter au texte traditionnellement traité par les SRI. Le texte lui-même s'est transformé, cette mutation étant soutenue par l'expansion des réseaux sociaux et les contenus générés par les utilisateurs (UGC - *User Generated Content*). Les plateformes de *microblogging* telles que Twitter [1] sont l'exemple type de l'évolution des contenus textuels : le texte est limité à 140 caractères [2], les utilisateurs écrivent principalement en langage SMS, de nouvelles informations telles que les *hashtags* sont présentes dans le texte pour faciliter le suivi des sujets discutés, etc. Parallèlement, la quantité d'information produite et indexée peut donner le tournis. En avril 2017, on comptait ainsi 3,62 milliards d'internautes, produisant chaque minute 160 millions de mails, 450 000 tweets ou encore plus de 3300 articles de blogs [3]. Le web indexé était estimé à au moins 4,5 milliards de pages [4], et Google annonçait 3,5 millions de requêtes traitées par minute.

Côté **requêtes**, la requête booléenne des débuts de la RI a laissé place dans les années 2000 à des requêtes courtes formées de 2-3 mots-clés dont les moteurs de recherche Web devaient se satisfaire (Jansen et al., 2000). Aujourd'hui, l'augmentation du nombre de recherches sur téléphone ou tablette mobile voit l'avènement des requêtes en langage naturel, souvent exprimées à voix haute : la recherche conversationnelle (*conversational search*), telle que définie par Google en 2013, fait désormais partie de notre quotidien [5].

Côté **modèles et présentation des résultats**, les moteurs de recherche Web, après avoir longtemps proposé aux utilisateurs les fameux « 10 liens bleus » (*10 blue links*) en réponse à leur requête, incluent maintenant dans leurs pages de résultats des images, des vidéos ou encore des actualités (Haas et al., 2011). Lorsque la requête est une entité [6], les informations liées peuvent aussi être présentées dans un cadre séparé (voir Figure 1). L'utilisateur est placé au centre de la recherche, son contexte (et son profil lorsque disponible) permettant de fournir des résultats personnalisés (Tamine-Lechani et al., 2009). L'idée n'est plus de restituer des documents relatifs à une requête, mais de donner directement à l'utilisateur un aperçu global de l'information liée à **son** besoin.

C'est autour de cette dernière idée que la recherche d'information agrégée a été définie dès 2008, avec pour but de chercher et d'assembler dans une seule interface de l'information utile provenant d'une ou plusieurs sources (Arguello, 2017; Murdock and Lalmas, 2008). La quantité d'information disponible étant immense, le but est d'en faire le tri et de présenter à l'utilisateur un résultat « résumé » de son besoin. Les limitations de la traditionnelle liste de documents en réponse à une requête s'en trouvent largement atténuées : l'information pertinente n'est plus dispersée dans plusieurs documents, et le résultat présenté se focalise sur le besoin utilisateur (Kopliku et al., 2014).

La recherche d'information agrégée peut être décrite selon le schéma de la figure 2 (Kopliku et al., 2014). La requête est traitée et envoyée à une ou plusieurs sources d'information. Chaque source renvoie un ensemble de granules documentaires, qui seront sélectionnés et assemblés pour former la réponse finale.

La notion de source introduite dans le schéma est importante. Nous la définissons comme correspondant à un moteur de recherche reposant sur au moins une collection et un algorithme de recherche. Un granule documentaire (*information nugget* (Goecks, 2002)) est un

1. http://www.twitter.com, dernier accès mars 2018.
2. Cette limitation a été repoussée à 280 caractères fin 2017.
3. Source : http://www.internetlivestats.com/, accédé le 24/04/2017.
4. Source : http://www.worldwidewebsize.com/, accédé le 24/04/2017.
5. Source:http://searchengineland.com/googles-impressive-conversational-search-goes-live-on-chrome-160445, accédé le 24/04/2017.
6. Une entité est une « chose » qui peut être clairement identifiée (Chen, 1976). Nous reviendrons plus en détail sur cette définition dans le chapitre 3.

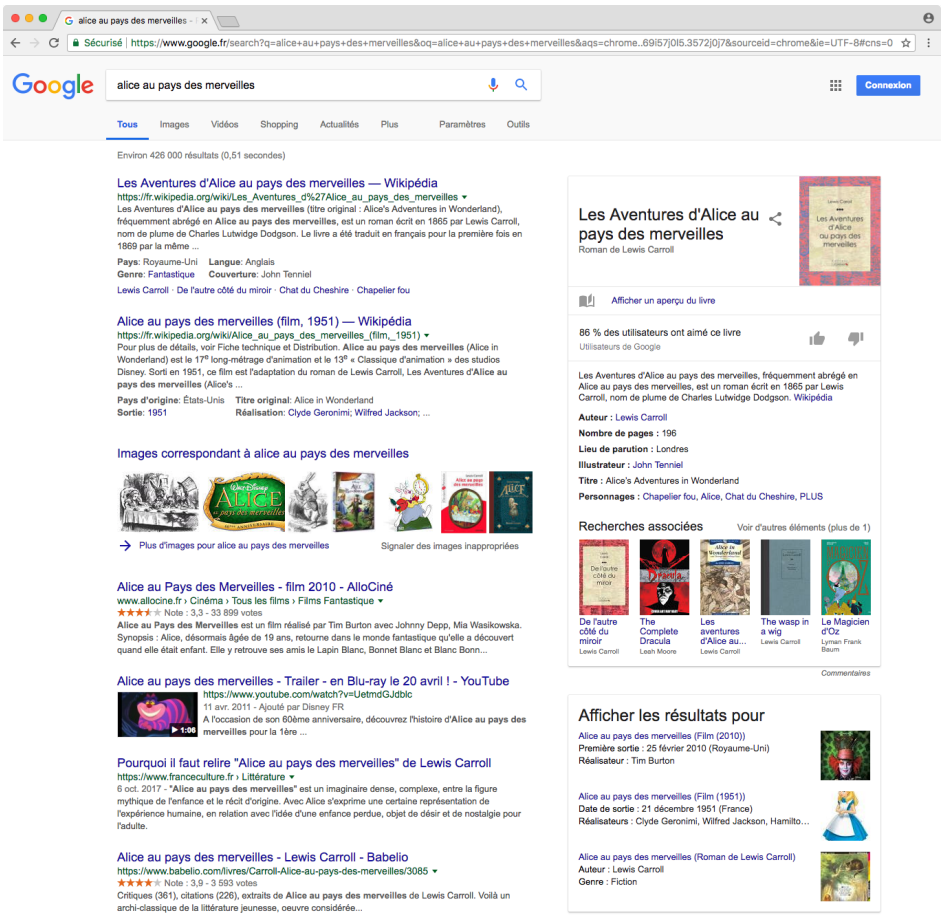


FIGURE 1 – Résultats d'une recherche Google effectuée le 20/03/2018 avec la requête `Alice au pays des merveilles`. Outre la présence d'images et d'une vidéo, on note un cadre à droite donnant des informations sur l'entité `Alice au pays des merveilles` et les entités liées (auteur, illustrateur, personnages, recherches liées). Ces informations sont affichées à l'aide du *Google Knowledge Graph* (Singhal, 2012).

contenu d'une granularité et d'un format multimédia donnés. De manière plus détaillée, nous identifions trois parties principales composant le processus de recherche :

— le ***dispatching* de la requête** est l'étape précédant la recherche proprement dite. Il s'agit d'interpréter correctement la requête, de cerner l'intention utilisateur, de reformuler éventuellement le besoin et de décider quelles sources d'information interroger.

— la **recherche de granules documentaires** se charge d'identifier l'information potentiellement pertinente. Chaque source, grâce à son algorithme de recherche, va renvoyer un ensemble de granules (avec éventuellement un score de pertinence associé). Il est possible d'obtenir en résultat des documents entiers, des parties de documents, ou encore des contenus multimédia issus de moteurs de recherche verticaux.

— l'**agrégation des résultats** cherche à assembler les différents granules documentaires afin de former le résultat final. Ce dernier, de façon idéale, devra refléter la diversité des résultats, répondre de façon exhaustive à la requête, et ne pas contenir de résultats redondants. Pour ce faire, différentes actions pourront être menées sur les granules : tri, regroupement, découpage en granules plus petits, extraction d'information, etc.

De nombreux thèmes de la RI ou connexes à la RI relèvent de la recherche d'information agrégée : parmi eux nous pouvons citer la génération de langage naturel (NLG-*Natural Language Generation*) (Paris et al., 2010), le question-réponse (Moriceau and Tannier, 2010), le résumé de documents ou le résumé multi-document (Goldstein et al., 2000), la recherche fédérée (*federated search*) aussi appelée recherche d'information distribuée (Callan, 2000), la

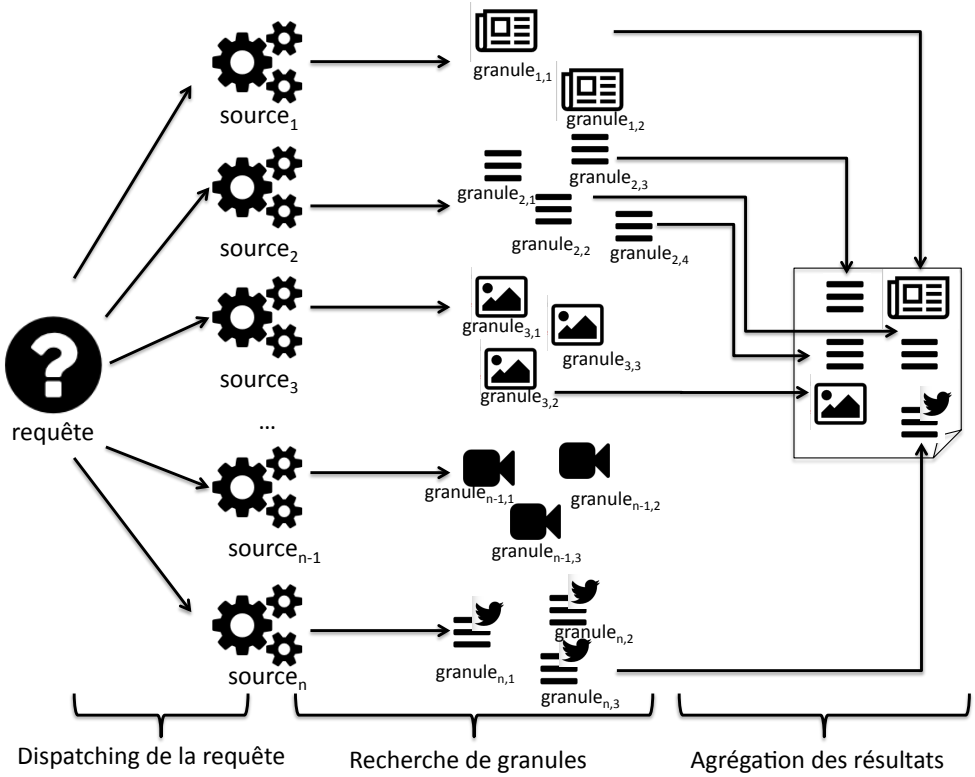


FIGURE 2 – Différentes parties de la recherche d'information agrégée (Kopliku et al., 2014).

recherche agrégée inter-verticale[7] (Arguello et al., 2012; Kopliku et al., 2011c), ou encore la recherche agrégée relationnelle (Kopliku et al., 2011e). Certains domaines ou applications spécifiques proposent également des approches entrant dans le cadre de la RI agrégée. Par exemple :

- la recherche académique agrège des informations autour des chercheurs (publications, conférences, thématiques de recherche, co-auteurs, indicateurs bibliographiques (voir des sites comme `http://academic.research.microsoft.com` ou `http://scholar.google.com/citations`)),
- les agrégateurs de nouvelles cherchent à regrouper les news en histoires (Hennig and Wurst, 2006; Hong et al., 2011),
- la recherche d'information géographique positionne une entité géographique sur une carte, en complétant l'information par une adresse, un numéro de téléphone, des images, ou encore des commentaires utilisateurs (Kennedy and Naaman, 2008; Vallet and Zaragoza, 2008).

Orientations des travaux

Les problématiques liées à la recherche d'information agrégée sont multiples et font l'objet de recherches intensives, en témoigne l'abondance de littérature sur le sujet dans la communauté. Nos recherches se sont plus particulièrement orientées vers :

- la **recherche de granules d'information**. Cet axe porte sur des questions de RI *adhoc* classique relatives à la sélection de granules pertinents répondant à une requête. Nous avons considéré trois sources d'information spécifiques :

7. De nombreux travaux de la littérature regroupent sous le même terme recherche agrégée inter-verticale et recherche agrégée (Arguello, 2017; Lalmas, 2011). Nous préférons garder la distinction dans les travaux présentés dans ce mémoire.

1. les collections de **documents semi-structurés de type XML**, afin de retrouver des **granules textuels ou image**. La structure inhérente aux documents de ce type permet de se focaliser sur le besoin utilisateur et d'identifier des granules documentaires répondant de façon exhaustive et spécifique au besoin. Nous avons plus particulièrement étudié l'apport des méthodes de propagation de la pertinence dans l'arborescence des documents, ainsi que l'intérêt de la structure, tant au niveau de la requête que pour la recherche de contenus textuels ou images.
2. les plateformes de *microblogging* de type Twitter. Les microblogs publiés sur ces plateformes impliquent des traitements spécifiques liés à leur brièveté et au langage utilisé, ainsi que le traitement temps-réel des informations. Nous nous sommes intéressés aux différentes caractéristiques pouvant être utilisées afin de décrire la **pertinence des microblogs**, ainsi qu'à la proposition d'algorithmes temps-réel.
3. le Web, pour des problématiques liées aux **requêtes de type entité**. Plus précisément, nous nous sommes intéressés à la recherche de **relations** autour d'une entité (relations avec une autre entité ou des attributs la définissant), ainsi qu'à la recherche de **documents « frais »** lorsque l'information liée à l'entité varie beaucoup dans le temps (c'est le cas par exemple pour l'actualité autour de personnes célèbres ou bien encore pour l'information publiée lors de catastrophes naturelles).

— l'**agrégation des résultats**. Nous avons considéré le problème de l'**agrégation des résultats autour de requêtes de type entité**. Plus particulièrement, nous avons travaillé sur l'**agrégation des relations liées à cette entité**, ainsi que sur le **résumé temporel** d'informations provenant de documents pertinents autour de l'entité.

J'ai également mené des travaux transverses dans le cadre de l'**évaluation**. La communauté RI a une forte tradition d'évaluation et d'utilisation de collections de test, en témoignent les succès des campagnes d'évaluation TREC [8] ou CLEF [9]. Lorsque ces campagnes ne fournissent pas de collections de test appropriées pour une problématique donnée, ces collections doivent être construites. Nous avons participé à de tels travaux de montage de collections de test dans le cadre du projet européen Quaero. Nous avons également réfléchi à des protocoles d'évaluation spécifiques, pour la recherche d'information agrégée inter-verticale ou le filtrage temps-réel de microblogs.

La figure 3 résume les différents axes investis et contributions au domaine en les plaçant au sein du cadre général de la recherche d'information agrégée. J'ai fait le choix dans ce mémoire d'inscrire dans ce cadre tous les travaux menés depuis 2005, année de soutenance de ma thèse dans le domaine de la recherche d'information structurée (Sauvagnat, 2005). Même si certaines de mes recherches n'ont pas été directement dirigées par les problématiques liées à la RI agrégée, elles s'y inscrivent rétrospectivement aisément.

Présentation du plan

Ce mémoire s'articule de la façon suivante :

— Les chapitres 1 à 3 s'intéressent à la sélection des unités d'informations nécessaires à l'agrégation dans des contextes particuliers. La chapitre 1 se focalise sur les documents semi-structurés de type XML. Le chapitre 2 présente des approches pour la recherche de microblogs en temps-réel. Le chapitre 3 est consacré aux requêtes de type entité et à la recherche de relations liées ainsi que de documents frais.

— Le chapitre 4 s'intéresse à l'agrégation autour de requêtes entité, sous forme de construction de résultats tabulaires, ou encore de résumé temporel.

8. http://trec.nist.gov, dernier accès mars 2018
9. http://www.clef-initiative.eu/, dernier accès mars 2018

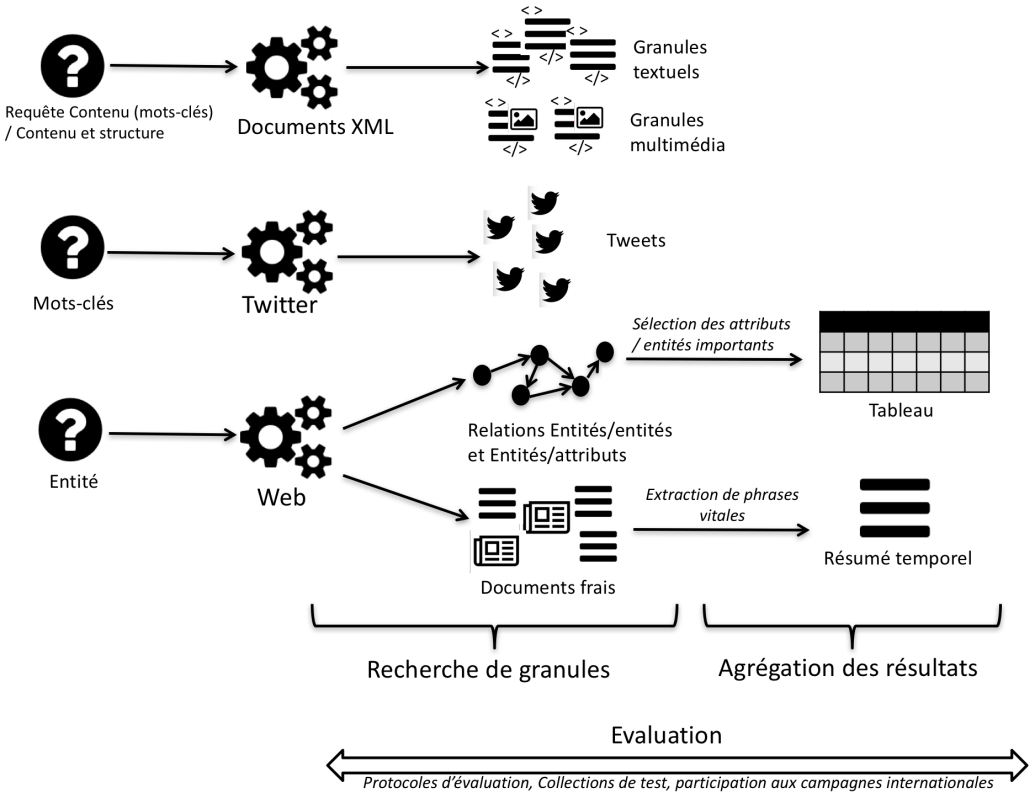


FIGURE 3 – Orientation de mes travaux au sein du cadre général de la recherche d'information agrégée.

— Le chapitre 5 s'intéresse à des problématiques d'évaluation et à l'élaboration de collections de test.

— La conclusion présente un bilan de tous ces travaux et précède un ensemble de perspectives sur des recherches futures.

Afin de ne pas surcharger la description des expérimentations effectuées pour valider nos travaux, l'annexe A regroupe le détail de toutes les collections de tests issues de campagnes d'évaluation internationales que nous avons utilisées (tâches de recherche, corpus et nombre de requêtes). Enfin, en annexe B se trouve mon CV académique, synthétisant et résumant les projets, encadrements, et publications auxquels j'ai participé, ainsi que les différentes responsabilités assurées ces dernières années.

CHAPITRE 1

Recherche de granules d'information : documents XML et information structurelle

Off with their heads!

— Queen of Hearts

Sommaire

1.1 État de l'art et problématiques ciblées

1.1.1 Introduction

Depuis le début des années 1990, les documents textuels « plats » ne contenant que du texte ont évolué, en s'enrichissant à la fois d'informations structurelles et multimédia. Cette évolution, accélérée par l'expansion du Web, a été permise grâce au développement de nombreux formats structurés ou semi-structurés, tels que le format HTML (*Hypertext Markup Language*) ou encore le format XML (*eXtensible Markup Language*) (W3C, 1998a). Les documents XML peuvent être de deux types différents :

— les documents orientés données possèdent une structure très régulière et l'ordre des informations qu'ils contiennent importe peu. Ces documents sont utilisés principalement à des fins de transfert d'information ou de sauvegarde de données qui auraient pu être stockées dans des bases de données relationnelles.

— les documents orientés texte ont des structures plus flexibles et des contenus hétérogènes. Le texte contenu doit être lu dans l'ordre du document pour être compréhensible. Ces documents sont utilisés pour stocker des articles scientifiques, des livres, de la documentation, etc.

Les travaux synthétisés dans ce chapitre s'intéressent particulièrement à la recherche d'information dans des documents XML semi-structurés, possédant une structure flexible et des contenus textuels hétérogènes. Par abus de langage, on parlera de RI structurée.

Dans ce contexte, le texte, structuré grâce à la notion de balise, devient directement fractionnable en granules documentaires, pouvant répondre de façon plus directe et précise au besoin utilisateur. On trouvera un exemple de document XML sur la figure 1.1. Chaque document XML peut-être représenté sous forme d'arbre (arbre DOM (W3C, 1998b)), comme le montre la figure 1.2. Dans un arbre XML, l'information textuelle est stockée au niveau des nœuds feuille, et chaque nœud de type élément peut être considéré comme un granule documentaire susceptible d'être renvoyé à l'utilisateur. Ces nœuds possèdent un label (un nom de balise) décrivant leur contenu.

Afin de tirer parti au mieux des informations de structure des documents, les techniques de RI se sont adaptées, et ce à plusieurs niveaux du processus de recherche :

— au niveau de l'indexation, afin de stocker en plus des mots-clés les informations de structure associées ;

— au niveau de l'interrogation, afin de permettre à l'utilisateur de formuler ses besoins à la fois en termes de contenu et de structure (on parle alors de requête « contenu » et de requête « contenu et structure ») ;

— au niveau de la recherche, où la structure permet d'identifier des granules documentaires répondant de façon spécifique et exhaustive au besoin utilisateur.

La structure des documents permet ainsi de se focaliser sur le besoin utilisateur, en renvoyant des granules documentaires (c'est-à-dire des éléments) de taille variable. Les documents ne sont plus considérés dans leur globalité.

1.1.2 État de l'art

De très nombreux travaux ont été menés dans les années 2000 en *RI Structurée (RIS)*. Ils ont été soutenus dès 2002 par la campagne d'évaluation INEX [1] (*INitiative for the Evaluation of XML Retrieval*), qui a proposé de nombreuses tâches de recherche permettant l'évaluation reproductible des approches. On trouvera dans (Lalmas, 2009; Pinel-Sauvagnat and Chrisment, 2008) des états de l'art complets sur l'état des lieux en RI XML au milieu des années 2000. Voici de façon synthétique ce que l'on peut en retenir, au niveau de l'interrogation, de la recherche de granules, et de l'évaluation.

1. http://inex.mmci.uni-saarland.de/, dernier accès en mars 2018.

```
<?xml version="1.0" encoding="UTF-8"?>
<destination language="eng">
   <name>Wonderland</name>
   <general>
      <introduction>Fantasy world populated by peculiar, anthropomorphic creatures.</p>
      </introduction>
      <timezone>
          <gmt_utc>N/A</gmt_utc>
          Time is disturbed, to such an extent that it lacks to the White Rabbit always in a hurry, or it
is fixed for the Hatter who was punished by Time by eternally standing still at 6 pm (tea time).
      </timezone>
   </general>
   <environment>
      <area_sqkm>Not known</area_sqkm>
      <population>20 main characters, including the Cheshire Cat and the Hatter</population>
      <fauna>
       <desc>The animals of Wonderland are of particular interest, for Alice's relation to them shifts
constantly because, as Lovell-Smith states, Alice's size-changes continually reposition her in the food
chain, serving as a way to make her acutely aware of the "eat or be eaten" attitude that permeates
Wonderland.</desc>
        <image filename="/images/BN603_21.jpg">
         <caption>Alice trying to play croquet with a Flamingo.</caption>
         <illustrator>John Tenniel</illustrator>
        </image>
      </fauna>
      <flora>
        <desc> Some  mushrooms. Flora is a mix of virgin and continental forests.</desc>
        <image filename="/images/BN4567_13.jpg">
         <caption>The Caterpillar using a hookah on a mushroom</caption>
         <illustrator>John Tenniel</illustrator>
        </image>
      </flora>
   </environment>
   <culture>
       <history>  Wonderland is a place of contestation by means of absurdity of the real world pre-
established order.
       </history>
       <art> … </art>
       <music>… </music>
   </culture>
<!--- ... !-->
 </destination>
```

FIGURE 1.1 – Exemple de document XML

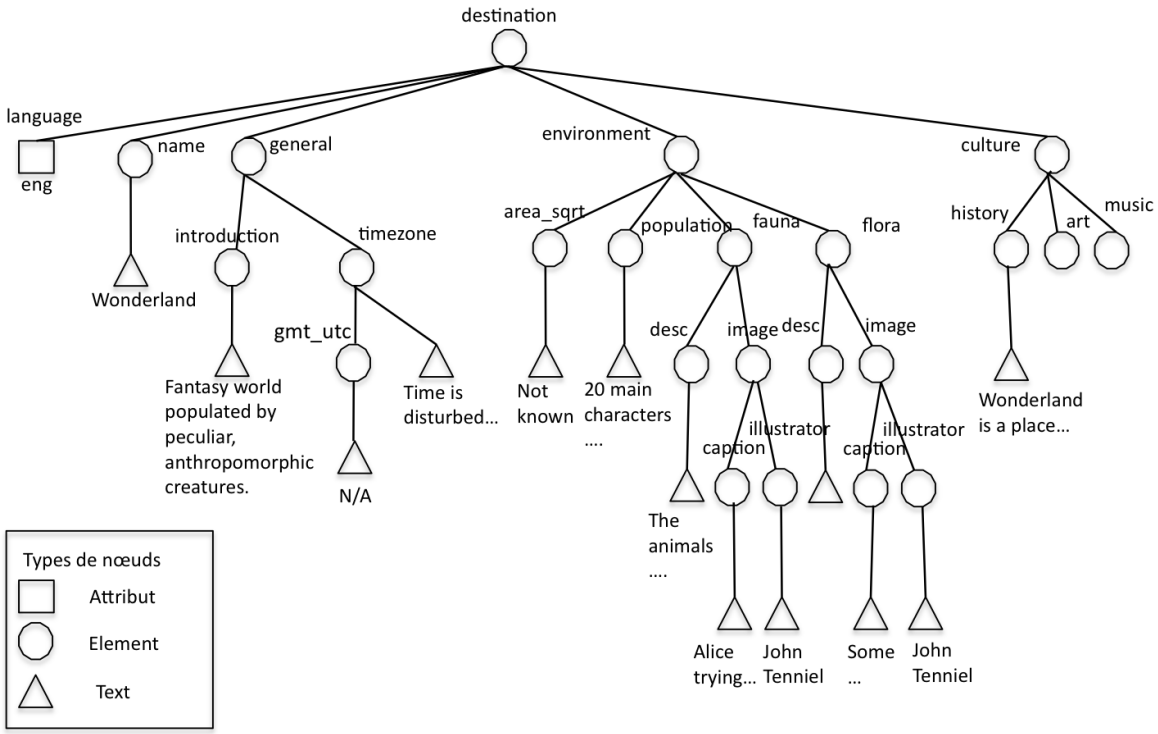


FIGURE 1.2 – Arbre DOM du document XML de la figure 1.1

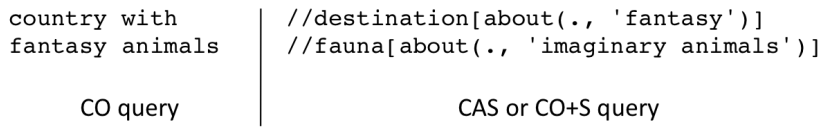


FIGURE 1.3 – Exemples de requêtes Content-Only et Content-and-Structure

1.1.2.1 Interrogation

Les premiers langages proposés pour l'interrogation de documents XML l'ont été par la communauté des bases de données. XQuery, finalement normalisé en 2010 (Boag et al., 2010), et un peu plus tard XQuery Full-text (Case et al., 2011), intégrant quelques primitives liées à la recherche de texte, en sont les exemples les plus remarquables. De son côté, la communauté RI a proposé le langage NEXI, qui a été utilisé de façon intensive durant la campagne d'évaluation INEX (Trotman and Sigurbjörnsson, 2004). NEXI a été conçu comme un sous-ensemble extensible d'XPath interprétable de manière vague. Il définit deux types de requêtes principales : les requêtes orientées contenu (*Content Only* - CO) composées de simples mots-clés, et les requêtes orientées contenu et structure (*Content and Structure* - CAS, ou *Content+ Structure* - *CO+S*) possédant des conditions de structure associées à des mots-clés. La figure 1.3 donne deux exemples de requêtes.

1.1.2.2 Recherche

L'appariement requête/granules documentaires est au cœur de la RI XML. La plupart des approches de l'état de l'art ont étendu des modèles de RI existants pour prendre en compte la structure des documents. On peut ainsi citer des extensions du modèle vectoriel (Mass and Mandelbrod, 2003; Schlieder and Meuss, 2002) ou encore des modèles probabilistes (modèle d'inférence probabiliste (Govert et al., 2002), modèles de langue (Kamps et al., 2004), réseaux bayésiens (Piwowarski et al., 2002)). Une autre façon de classer les approches est liée à la façon dont elles gèrent les termes par rapport à la structure des documents :

— certaines propagent les termes dans l'arbre du document et les agrègent sur chaque élément afin d'évaluer un score de pertinence pour chaque élément : on parle alors d'approches orientées agrégation (Tsikrika, 2009),

— d'autres évaluent un score au niveau des nœuds feuille et propagent ce score dans l'arbre afin d'évaluer la pertinence des nœuds internes, c'est-à-dire des éléments : on parle alors d'approches orientées propagation de la pertinence (Pinel-Sauvagnat, 2009, 2017).

1.1.2.3 Évaluation

L'évaluation de la recherche dans des documents semi-structurés est très liée à la campagne d'évaluation INEX, lancée pour la première fois en 2002, et dont la dernière édition a eu lieu en 2014 en tant que « *lab* » de la campagne CLEF (*Conference and Labs of the Evaluation Forum*). La campagne a fourni de nombreuses collections d'évaluation, et a permis une réflexion poussée sur les mesures d'évaluation à adopter (Kazai, 2009). La prise en compte de la spécificité et de l'exhaustivité des éléments retournés dans les mesures d'évaluation, avec notamment le phénomène des *near-misses* a été un problème majeur : un système renvoyant une section contenant un paragraphe pertinent ne doit pas être complètement pénalisé s'il n'a pas précisément renvoyé le paragraphe. De plus, un système A renvoyant une section pertinente et un de ses paragraphes également pertinent ne doit pas être évalué de la même façon qu'un système B renvoyant deux éléments pertinents non imbriqués (problème de l'*overlap* (Kazai et al., 2004)). La mesure XCG a finalement été retenue au bout de quelques années pour l'évaluation de la recherche ad-hoc (Kazai and Lalmas, 2006).

1.1.3 Problématiques ciblées

Comme nous l'avons vu, la recherche d'information structurée s'est à ses débuts focalisée sur les primitives de base de la recherche d'information, à savoir trouver des langages de requêtes appropriés, des modèles de recherche adaptés et des méthodologies d'évaluation associées. Après ces débuts nécessaires, la communauté s'est ensuite attaquée à des problématiques plus spécifiques, comme la gestion des structures hétérogènes (Malik et al., 2004), la classification et le *clustering* de documents XML (Denoyer and Gallinari, 2008), ou l'utilisation de documents XML pour la recherche d'entités (Demartini et al., 2009) ou d'éléments multimédia (Westerveld and van Zwol, 2007b).

De notre côté, nous avons considéré les pistes de recherche suivantes :

— comment traiter les requêtes très fortement structurées ? Ces requêtes, issues à l'origine de la communauté des bases de données et s'appliquant sur des documents orientés données, c'est-à-dire destinés à stocker des données au sens base de données, peuvent cependant elles-aussi bénéficier d'un traitement orienté RI afin de classer les résultats de recherche.

— le processus de reformulation de requête a montré son intérêt en recherche d'information. Comment peut-on l'appliquer en RI structurée ? L'ajout de structure à une requête qui en est à la base dépourvue a-t-il un intérêt ?

— si l'utilisation de la structure des documents a montré son intérêt pour la recherche d'unités textuelles (c'est-à-dire d'unités contenant du texte), qu'en est-il des éléments multimédia inclus dans les documents ?

Les travaux initiés pendant ma thèse (Sauvagnat, 2005) ont permis le développement d'un moteur de recherche, XFIRM (*XML Flexible Information Retrieval Model*), basé sur une méthode de propagation de la pertinence (Pinel-Sauvagnat, 2009).

Ce moteur de recherche a servi de base aux travaux menés par la suite, dont l'objectif principal a été d'améliorer la recherche d'unités textuelles (que ce soit dans le cadre de requêtes faiblement structurées ou très structurées, ou encore avec l'aide de techniques de reformulation de requêtes) et de proposer des approches pour la recherche d'unités multimédia.

1.2 Contributions au domaine de recherche

Nos recherches sont présentées ci-dessous selon deux axes : (i) la recherche de granules textuels, et (ii) la recherche de granules multimédia. Dans les deux cas, nous nous sommes attachés à intégrer l'information de structure des documents dans nos modèles de recherche. Des évaluations systématiques des algorithmes proposés ont été effectuées grâce aux collections de test fournies par les campagnes d'évaluation INEX et CLEF.

1.2.1 Recherche de granules textuels

Comme nous l'avons vu précédemment, la recherche de granules textuels (c'est-à-dire d'éléments XML) se fait en réponse à des requêtes constituées de simples mots-clés (requêtes CO) ou encore de requêtes constituées de mots-clés et de conditions de structure (requêtes CO+S ou CAS). Nos travaux ont proposé des modèles de recherche permettant de traiter les deux types de requêtes, et se sont également intéressés à la reformulation de requête par l'ajout de structure.

1.2.1.1 Cas de requêtes peu ou pas structurées

Mes premiers travaux se sont focalisés sur le modèle de pondération proposé dans le modèle XFIRM. L'idée a été d'intégrer le contexte des nœuds à divers niveaux de granularité, c'est-à-dire d'intégrer la pertinence de leurs nœuds ancêtres et bien sûr du document les contenant

pour leur pondération. L'intuition était qu'un nœud dans un nœud/document très pertinent était plus probablement pertinent qu'un nœud dans un nœud/document peu pertinent. Nous avons proposé plusieurs solutions pour intégrer ce contexte, la plus notable étant la retro-propagation de la pertinence, permettant de propager vers un nœud la pertinence de ses nœuds ancêtres (un nœud voit donc sa pertinence évaluée grâce à une propagation du bas vers le haut de la pertinence de ses nœuds feuille, et du haut vers le bas de la pertinence de ses ancêtres) (Pinel-Sauvagnat and Boughanem, 2006; Sauvagnat and Boughanem, 2006; Sauvagnat et al., 2005, 2006c). Les expérimentations menées sur les collections INEX 2003 à 2005 ont montré l'impact positif de l'intégration du contexte dans l'évaluation de la pertinence des nœuds dans le cadre d'une évaluation *Thorough*[2], c'est-à-dire lorsque le but du modèle est de renvoyer tous les nœuds pertinents (même s'ils sont imbriqués les uns dans les autres).

Le modèle XFIRM, bien que présenté ici pour répondre à des requêtes CO - *Content-Only* est également capable de traiter des requêtes contenant des conditions de structure (Sauvagnat et al., 2006b). Nous avons étudié plus en détail l'impact de cette structure dans les requêtes : aide-t-elle à la recherche ? Nos résultats ont montré son intérêt dans le cadre de notre modèle (Pinel-Sauvagnat and Boughanem, 2007; Sauvagnat et al., 2006a). Il n'y a cependant pas de clair consensus dans la littérature sur l'intérêt d'exprimer des conditions de structure dans les requêtes (Schlieder and Meuss, 2002; Trotman, 2009; Trotman and Lalmas, 2006). Nos travaux suivants ont donc continué à creuser la question, soit en nous intéressant particulièrement aux requêtes très structurées, soit en injectant de (nouvelles) structures dans les requêtes, dans le cadre d'un mécanisme de réinjection de pertinence.

1.2.1.2 Cas des requêtes très structurées

Lorsqu'une requête est très structurée, une idée intuitive est de comparer la structure des documents avec celle de la requête pour aider à renvoyer les nœuds les plus pertinents, et ainsi ajouter la notion de pertinence structurelle à celle de pertinence de contenu. Lorsqu'une requête contient des conditions de structure, elle détermine également un label d'élément cible, c'est-à-dire le type de nœud espéré en réponse à la requête. La figure 1.4 donne deux exemples de requêtes structurées et leur représentation sous forme d'arbre. Le nœud en noir porte le label de l'élément cible. Les autres éléments sont appelés éléments supports.

Dans ce contexte, nous avons investigué deux pistes principales :

— Tout comme les documents XML, nous avons vu que les requêtes structurées peuvent être représentées sous forme d'arbres. À notre connaissance, peu de travaux se sont cependant intéressés directement à l'utilisation de la théorie des graphes pour la recherche (Tahraoui et al., 2013). Pour évaluer l'intérêt de ce type d'approche, nous avons choisi d'adapter l'algorithme de la distance d'édition d'arbres à notre problématique (Laitang and Pinel-Sauvagnat, 2011; Laitang et al., 2011, 2012a,b, 2013a; Le and Pinel-Sauvagnat, 2010)

— Nous avons proposé d'évaluer la pertinence structurelle des documents en utilisant des modèles de langues (Laitang et al., 2013b). Notre approche considère que le modèle de langue « structurel » d'un document est formé des liens structurels entre labels.

Quelle que soit la piste considérée, l'appariement arbre de la requête / arbre des documents ne peut se faire sans prendre en compte les conditions de contenu de la requête. Nous évaluons pour ce faire un score de pertinence de contenu pour tous les nœuds n ayant pour label le label de l'élément cible. De façon synthétique, ce score est calculé en fonction de la pertinence des nœuds feuille qu'il contient et de la pertinence de son contexte, selon une méthode de propagation de la pertinence (Laitang and Pinel-Sauvagnat, 2011). Pour chaque nœud n de score de contenu non nul, on extrait alors un ou plusieurs sous-arbres représentatifs qui seront utilisés pour apparier les conditions de structures.

2. Ce terme est emprunté à la terminologie INEX.

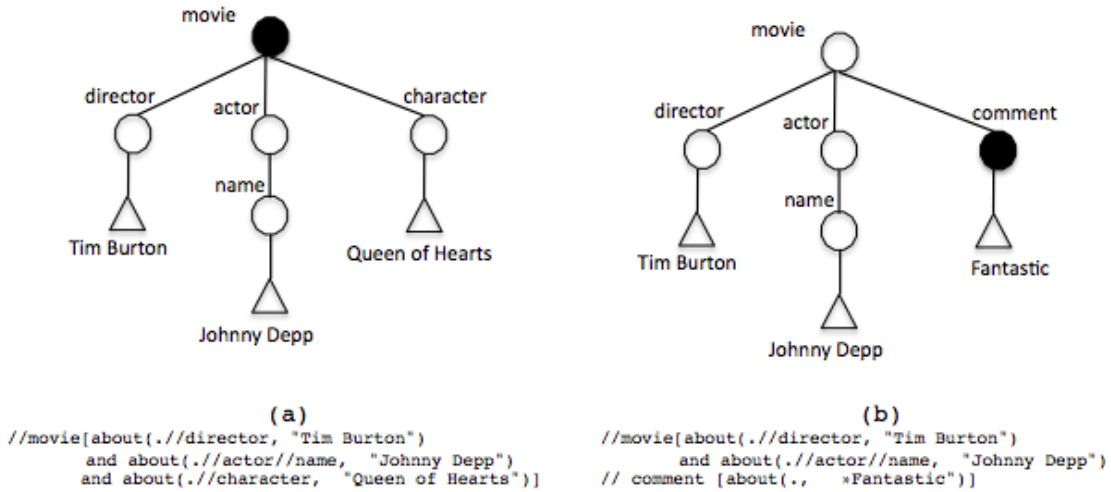


FIGURE 1.4 – Exemples de requêtes très structurées. L'élément en noir est l'élément cible. La requête (a) peut se traduire par `Je cherche un film ayant pour réalisateur « Tim Burton », pour acteur « Johnny Depp » et pour personnage la reine de coeur (« Queen of Hearts »)`. La requête (b) peut se traduire par `Je cherche un commentaire d'un film réalisé par « Tim Burton » et ayant pour acteur « Johnny Depp ». Ce commentaire doit contenir le terme « Fantastic »`.

La qualité de l'extraction de ces sous-arbres est donc primordiale. Là encore, nous avons considéré deux cas :

(i) toutes les contraintes de contenu de la requête sont concaténées pour évaluer le score de pertinence de contenu des nœuds n ayant pour label le label de l'élément cible, et nous considérons ensuite pour chaque n ayant un score non nul tous les sous-arbres enracinés par ses ancêtres (Laitang et al., 2012b). Le score structurel d'un nœud n est ensuite la moyenne du score des sous-arbres enracinés par ses ascendants.

(ii) les contraintes de contenu sont prises séparément et évaluées sur tous les nœuds feuille de la collection. À partir des nœuds n ayant un label correspondant à celui de l'élément cible, nous extrayons ensuite un sous-arbre minimal pour mesurer la similarité structurelle (Laitang et al., 2013a). Ce sous-arbre est composé de la concaténation de tous les chemins d'intérêt du document auquel appartient le nœud n. Pour chaque nœud feuille de score de contenu non nul, un chemin d'intérêt est un chemin partant du nœud le plus bas de la hiérarchie possédant un label appartenant à la requête jusqu'au noeud possédant le label de la racine de la requête. Cette opération revient donc à faire un élagage de l'arbre du document.

La figure 1.5 montre pour une même requête et un même document les deux cas précédemment cités à savoir : (i) l'extraction des sous-arbres enracinés par les ancêtres du nœud cible (ii) l'extraction du sous-arbre minimal correspondant.

Les expérimentations présentées dans (Laitang, 2013) montrent que la prise en compte des contraintes de contenu de façon séparée est importante dans le cas de collections orientées données (collection IMDB de la campagne INEX). L'impact est moindre dans le cas de collections orientées texte (collection IEEE de la campagne INEX). Le reste de cette section se focalise maintenant sur nos méthodes d'appariement structurel (par distance d'édition ou en utilisant les modèles de langue), indépendamment de la méthode d'extraction de sous-arbres utilisée. Les scores de contenu et de structure seront ensuite combinés pour renvoyer à l'utilisateur une liste de nœuds pertinents.

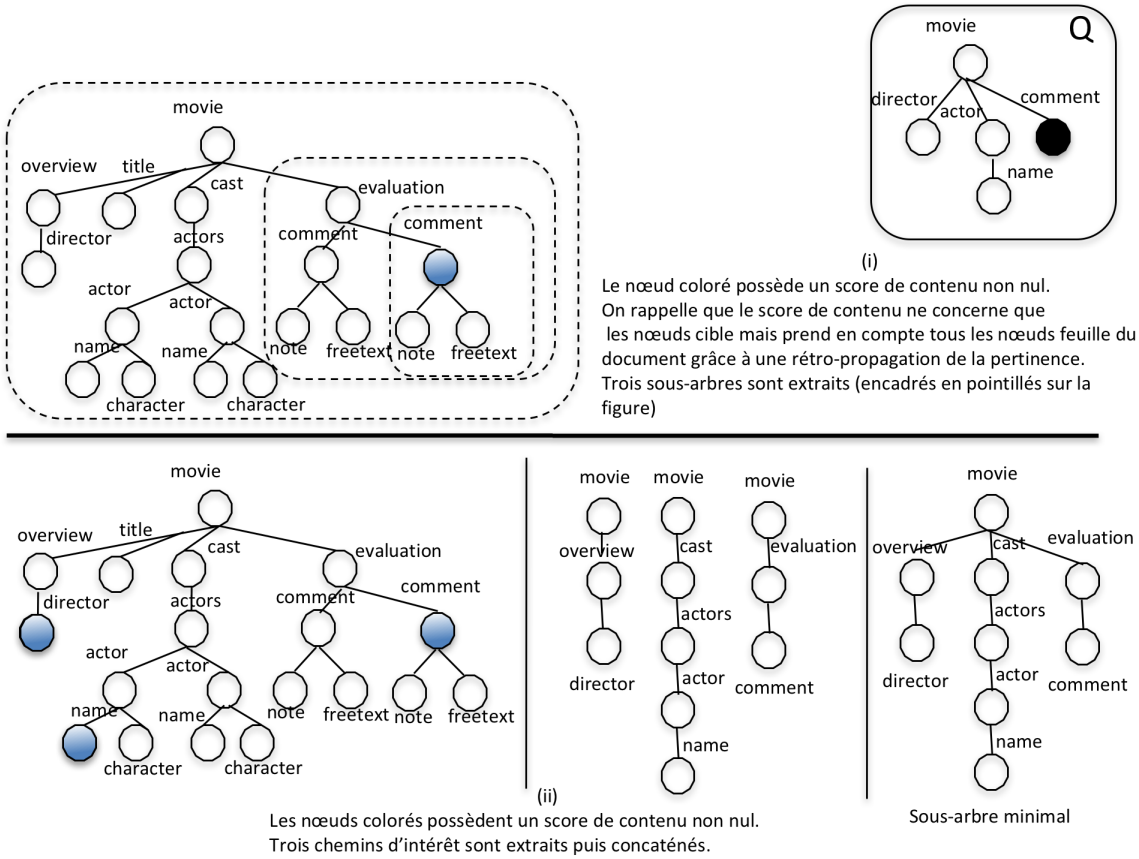


FIGURE 1.5 – Exemples d'extraction d'arbres pour l'appariement structurel : (i) sous-arbres enracinés par les ancêtres, ou (ii) sous-arbre minimal.

Utilisation de la distance d'édition On trouve dans la littérature trois grandes manières d'appréhender la structure dans un système de RI structurée au travers de l'utilisation de la théorie des graphes. Un premier groupe d'approches se base sur la relaxation, c'est-à-dire une réduction du niveau de restriction sur les liens ou les contraintes de structures (Alilaouar, 2007; Ben Aouicha et al., 2010; Saito and Morishita, 2006). Une autre famille d'approches utilise la fermeture floue (Damiani et al., 2001, 2000) : un ensemble d'arcs virtuels représentatifs du niveau hiérarchique est créé sur l'arbre du document. On trouve enfin quelques rares applications des algorithmes de la distance d'édition. Popovici et al. (2005) a par exemple utilisé la distance d'édition des chaînes de caractères de Levenshtein (1966) en ramenant le graphe sous la forme d'un chemin. Ce manque d'intérêt pour les algorithmes de la théorie des graphes peut être expliqué par une complexité qui telle quelle reste relativement élevée. L'ensemble des approches cherchent toujours d'une manière ou d'une autre à ramener le problème sur un espace réduit en traduisant les arbres. Cette réduction élimine malheureusement une partie importante des informations sur l'environnement des balises, environnement dont la prise en compte nous semble pouvoir apporter un réel gain de pertinence dans l'évaluation des éléments à retourner.

Estimer la similarité structurelle entre deux arbres (requête et document) revient à rechercher le degré d'isomorphisme entre les deux arbres. Parmi les quatre grandes familles d'appariement approximatif d'arbres (recherche d'un sous-graphe commun maximal, distance d'édition (Tai, 1979), distance d'alignement (Jiang et al., 1995) et inclusion d'arbres (Chen, 1998)), nous nous sommes intéressés à l'adaptation de l'algorithme de la distance d'édition à notre problème de RIS (Laitang and Pinel-Sauvagnat, 2011). L'algorithme de la distance d'édition se base sur la combinaison de trois opérations élémentaires que sont l'ajout, ou la suppression d'un nœud dans l'arbre source, et la substitution des labels des nœuds source et cible. La similarité se mesure par le nombre minimal d'opérations permettant de rendre le

graphe source isomorphe au graphe cible.

Parmi les algorithmes de la distance d'édition que nous avons explorés, notre choix s'est porté sur une stratégie de couverture optimale telle que définie par (Dulucq and Touzet, 2003) obtenue au travers des chemins lourds de (Klein, 1998). L'algorithme de la distance d'édition étant récursif, le nombre de sous-arbres stockés en mémoire durant la récursion et donc la complexité, dépend de la direction choisie pour appliquer les opérations. Le chemin lourd est formé par la suite des nœuds ayant le plus de descendants. On l'obtient donc en calculant le chemin de la racine jusqu'aux feuilles passant par les nœuds dont les sous-arbres enracinés possèdent la cardinalité la plus élevée. Sélectionner en permanence le nœud le plus éloigné de ce chemin permet de minimiser le nombre de sous-arbres en mémoire pendant la récursion. Nous nommons les chemins choisis pour la récursion les chemins **optimaux** (Laitang et al., 2011).

Notre algorithme de la distance d'édition est donc le suivant :

```
1  d(F, G, p_F, p_G) begin
2      if F = ∅ then
3          if G = ∅ then
4              return 0;
5          else
6              return d(∅, G - O_G.get(p_G)), p_F, inc(p_G)) + c_del (O_G.get(p_G));
7          end
8      end
9      if G = ∅ then
10         return d(F - O_F.get(p_F)), ∅, inc(p_F), p_G) + c_del (O_F.get(p_F));
11     end
12     a = d(F - O_F.get(p_F), G, inc(p_F), p_G) + c_del (O_F.get(p_F));
13     b = d(F, G - O_F.get(p_G), p_F, inc(p_G) + c_del (O_G.get(p_G));
14     c = d(T(O_F.get(p_F)) - O_F.get(p_F), T(O_G.get(p_G)) - O_G.get(p_G), inc(p_F), inc(p_G))
         + d(F - T(O_F.get(p_F)), G - T(O_G.get(p_G)), next(p_F), next(p_G)) + c_match
         (O_F.get(p_F), O_G.get(p_G));
15     return min(a, b, c);
16 end
```

Algorithme 1 : Distance d'édition par chemins optimaux

La forêt F correspond au sous-arbre extrait du document tandis que la forêt G est représentative de la requête. c_{del}, respectivement c_{match}, est la fonction de calcul de coût de suppression/insertion, respectivement de substitution. p_F et p_G sont les positions courantes dans les chemins optimaux de F et de G. Les opérations `next()` et `inc(`p_F`)` correspondent à un déplacement sur le chemin optimal utilisé par les opérations a, b et c de la distance d'édition d'arbres. Enfin, la fonction `get()` permet de retourner le nœud associé à la position sur le chemin.

Cet algorithme a été ensuite amélioré selon deux directions :

— Nous avons proposé de travailler sur un **résumé structurel** des arbres de documents (Laitang et al., 2011). Étant donné même les meilleurs algorithmes de distance d'édition ont une complexité en $O(n^3)$ (Dulucq and Touzet, 2003), réduire la taille des arbres en entrée est particulièrement intéressant. Nous avons proposé de réduire l'espace des nœuds en utilisant les règles de résumé de Dalamagas et al. (2006) [3]. D'une manière synthétique, ces règles consistent à supprimer les nœuds possédant des labels identiques et imbriqués les uns dans les autres, et enlever les duplicats sur les relations parent-enfant identiques.

3. Ces résumés ont seulement un intérêt dans le cas de l'extraction de sous-arbres multiples pour l'appariement, voir Figure 1.5, (i).

FIGURE 1.6 – Exemple d'une DTD et du graphe correspondant, extrait de (Laitang et al., 2013a).

— Dans notre domaine d'application, les arbres des documents sont généralement plus gros que les arbres requêtes, ce qui implique que les coûts de suppression doivent être diminués. De plus, en fonction des collections, certains nœuds peuvent être plus importants que d'autres et donc avoir des coûts associés plus importants. Nous avons proposé de nous **baser sur la DTD (*Document Type Definition*) des documents pour estimer les fonctions c_{del} et c_{match}**, plutôt que d'utiliser des valeurs empiriques comme nous l'avons fait dans un premier temps (Laitang et al., 2012b, 2013a). Nous avons représenté la DTD sous forme de graphe (voir Figure 1.6) et avons estimé les coûts c_{match} et c_{del} de la façon suivante :

  — pour évaluer le coût de substitution $c_{match}(n_1, n_2)$ d'un nœud n_1 par un nœud n_2 associés respectivement aux labels l_1 et l_2, nous cherchons le plus court chemin $sp()$ dans le graphe de la DTD en utilisant un algorithme de Floyd-Warshall (Floyd, 1962), permettant d'éviter les problèmes de cycles infinis.

$$c_{match}(n_1, n_2) = \frac{sp(l_1, l_2)}{max_{x \in DTD}(sp(l_1, l_x))} \tag{1.1}$$

  — De manière similaire, le coût de suppression est le plus grand coût parmi les coûts de substitution entre le nœud courant et tous les autres nœuds de la requête.

$$c_{del}(n_1) = max_{y \in DTD}\left(\frac{sp(l_1, l_y)}{max_{x \in DTD}(sp(l_1, l_x))}\right) \tag{1.2}$$

Nos résultats sur une collection orientée texte (INEX 2005 - IEEE) et une collection orientée données (INEX 2010 - IMDB) montrent que la distance d'édition permet d'améliorer les résultats d'une recherche par simples mots-clés et filtrage sur l'élément cible. De même, l'utilisation de la grammaire des documents (DTD) a montré son intérêt et semble être un bon indicateur de la mesure des coûts de la distance édition. Le résumé structurel permet quant à lui de gagner en temps d'exécution sans perdre en efficacité (facteur 2 ou 4 de temps d'exécution). Si l'on considère maintenant le type de la collection au regard des approches proposées, une extraction par sous-arbres multiples et résumé d'arbre semble plus efficace pour une collection orientée texte, alors qu'une extraction par sous-arbre minimal est plus intéressante pour une collection orientée données. Dans le dernier cas, la structure initiale de la requête doit être plus strictement conservée dans les arbres résultats (Laitang et al., 2013a). Le détail des expérimentations menées pourra être trouvé dans les articles précédemment cités ainsi que dans (Laitang, 2013).

Modèles de langues pour l'appariement Les modèles de langue en RI classent les documents selon la probabilité que le modèle de langue du document M_D génère la requête. Leur intérêt étant démontré en RI, ils ont été adaptés dans le cadre de la RIS (Hiemstra, 2003; Li and van der Weide, 2010; Ogilvie and Callan, 2005; Zhao and Callan, 2009). Les approches précédemment citées se concentrent cependant sur le contenu, en intégrant les contraintes de structure séparément.

Notre idée est de les étendre en considérant uniquement la structure. Au lieu de considérer les termes des documents et de la requête, nous considérons les arcs des arbres XML (Laitang

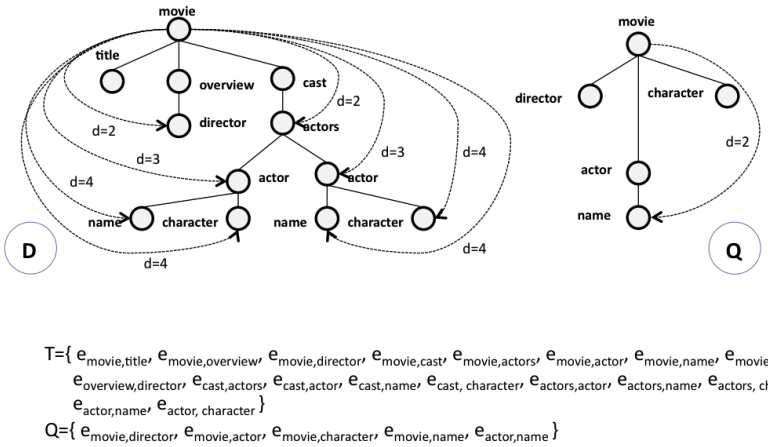


FIGURE 1.7 – Vocabulaire du document et de la requête, extrait de (Laitang et al., 2013b)

et al., 2013b). Nous considérons que les arbres des documents sont générés à partir d'un modèle de langue donné. Notre but est de trier les arbres T en fonction de leur probabilité de générer une requête Q composée de conditions de structure.

Formellement, un arbre $T = \{e_{i,j}\}$ est un ensemble d'arcs pondérés $e_{i,j}$, où i et j sont des labels de l'arbre. De façon similaire, une requête $Q = \{e_{i,j}\}$ est considérée comme un ensemble d'arcs pondérés représentatifs des contraintes structurelles. Plusieurs chemins de l'arbre peuvent être décrits par un arc $e_{i,j}$, car les nœuds les composant possèdent les mêmes labels. Ces chemins sont de la forme XPath `u//v`, avec u et v deux nœuds de T, avec $tag(u) = i$ et $tag(v) = j$, où $tag(x)$ est une fonction renvoyant le label d'un nœud. Chacun de ces chemins peut être vu comme une occurrence de $e_{i,j}$, dénotée par $e_{i,j}^{u \to v}$.

La figure 1.7 illustre l'extraction du vocabulaire d'un arbre document T et d'une requête Q. T est composé de 8 relations parent-enfant et de 10 arcs additionnels représentant les relations ancêtre-descendant. Par exemple l'arc $e_{movie,actor}$ a deux occurrences correspondant aux chemins XPath suivants : `/movie//actor[1]` and `/movie//actor[2]`.

Le vocabulaire de T et Q est ainsi composé de tous les arcs ancêtre-descendant extraits. Le poids de chaque arc dans T est la somme des poids de ses occurrences dans T :

$$w(e_{i,j}, T) = \sum_{\substack{u,v \in N^T / tag(u)=i \\ and\ tag(v)=j}} w(e_{i,j}^{u \to v}) \tag{1.3}$$

avec

$$w(e_{i,j}^{u \to v}) = exp(1 - d(u, v)) \tag{1.4}$$

d est la distance séparant u et v dans l'arbre. Cette pondération, avec $w(e_{i,j}^{u \to v}) \in [0, 1]$, a pour but de donner plus d'importance aux arcs formés par des noeuds proches.

Le poids final d'un arbre est ensuite évalué comme la somme de ses arcs pondérés (équation 1.5).

$$w(T) = \sum_{e_{i,j} \in T} w(e_{i,j}, T) \tag{1.5}$$

Par analogie avec les modèles de langue utilisés traditionnellement en RI, nous considérons que les arbres XML sont générés par un modèle de langue M_T. La requête est considérée comme un échantillon d'arcs générée à partir d'un modèle de langue d'arbre. Les arbres documents sont triés selon leur probabilité de générer la requête.

Afin d'évaluer la proximité structurelle $RSV(Q, M_T)$ entre un arbre T et une requête Q, nous calculons la probabilité $P(Q \mid M_T)$ que l'arbre issu du document génère la requête.

L'idée sous-jacente est d'évaluer la probabilité de chaque arc indépendamment (sous la forme d'uni-gramme). Nous avons ainsi :

$$P(Q \mid M_T) \quad = \quad \prod_{e_{i,j} \in Q} P(e_{i,j} \mid M_T)^{w(e_{ij},Q)} \tag{1.6}$$

Afin d'éviter les probabilités nulles liées aux arcs manquants (le même problème est observé sur les modèle de langues se basant sur le texte (Zhai and Lafferty, 2004)), nous avons ensuite expérimenté deux techniques de lissage différentes (*Jelinek-Mercer* (Jelinek and Mercer, 1980) et *Dirichlet* (MacKay and Bauman Peto, 1994)).

Les résultats obtenus sur la collection INEX *Datacentric* 2010 sont encourageants (Laitang et al., 2013b), avec sur cette collection des résultats légèrement supérieurs à ceux observés en utilisant la distance d'édition pour l'appariement de structures. Ils se situent au dessus des premiers participants à la campagne d'évaluation sur la mesure MAiP. Étant donnée la forte complexité de la distance d'édition d'arbre, cette approche semble donc prometteuse.

1.2.1.3 Reformulation de requêtes

La réinjection de pertinence est une stratégie couramment utilisée en RI pour améliorer les résultats de recherche. L'idée est d'extraire les termes des documents jugés pertinents et de les ajouter à la requête initiale afin d'affiner le besoin utilisateur (Rocchio, 1971; Salton and Buckley, 1997). Les spécificités de la RI XML en termes de requêtes, granules recherchés et estimation de la pertinence soulèvent de nouvelles problématiques pour les méthodes de réinjection de pertinence (Hlaoua et al., 2010) :

— tous les éléments jouent-ils le même rôle pour l'extraction des termes pertinents ? en d'autres termes, doit-on prendre en compte le label des éléments (`title`, `section`, `paragraph`,...) pour extraire les termes à ajouter à la requête initiale ?

— l'ajout de conditions de structure à la requête est-il d'intérêt ?

— lors de la réécriture de la requête, comment pondérer les nouveaux termes ? Comment combiner les termes aux éventuelles conditions de structure ?

Afin de répondre à ces problématiques, nous avons proposé (i) une approche orientée contenu pour enrichir la requête de termes pertinents, (ii) une approche orientée structure pour enrichir la requête avec des conditions de structure, et (iii) une approche combinée proposant de combiner les deux sources d'évidence, à savoir les termes et structures pertinents.

Approche orientée contenu De premières expérimentations menées sur le corpus d'INEX 2005 ont montré qu'une adaptation simple de l'algorithme de Rocchio sans pondération des termes n'était pas efficace sur des documents XML (Sauvagnat et al., 2005).

Afin d'extraire et de pondérer les termes à ajouter à la requête initiale, nous nous sommes inspirés du modèle probabiliste, déjà utilisé pour la reformulation de requêtes en RI classique (Robertson and Jones, 1976). Nous considérons la pertinence d'un terme t_j comme un événement probabiliste. La probabilité qu'un terme soit pertinent est définie par $P(t_j|E^r)$, où E^r est l'ensemble des éléments pertinents (Hlaoua et al., 2007e) :

$$P(t_j|E^r) = |p_{e_j}|/|E^r| \tag{1.7}$$

avec p_{e_j} l'ensemble des éléments pertinents contenant t_j.

Une sélection des termes selon cette seule probabilité conditionnelle traduit la dimension d'exhaustivité de la pertinence des éléments, mais n'est pas assez discriminante, car plusieurs termes peuvent avoir les mêmes scores alors qu'ils ne sont pas forcément de la même importance pour l'utilisateur. Afin de trouver des termes décrivant des éléments spécifiques [4], nous

4. Nous rappelons qu'en RI XML, la pertinence est décrite en termes d'exhaustivité et de spécificité.

avons proposé de prendre en compte la proximité des termes avec ceux de la requête initiale et défini la notion de contexte (Hlaoua et al., 2007c) :

$$context^{e_i}(t_j) = (distribution^{e_i}(q) - min^{e_i}(t_j))/distribution^{e_i}(q) \quad (1.8)$$

$$min^{e_i}(t_j) = min_{t_j \neq t_k}|(position^{e_i}(t_j) - position^{e_i}(t_k)| \quad (1.9)$$

$$distribution^{e_i}(q) = length(e_i)/occurrences^{e_i}(q) \quad (1.10)$$

où $distribution^{e_i}(q)$ est la distribution de tous les termes de la requête dans l'élément e_i avec $length(e_i)$ la taille de l'élément e_i et $occurrences^{e_i}(q)$ le nombre d'occurrences des termes de la requête q dans l'élément e_i. $min^{e_i}(t_j)$ est la différence minimale de positions entre n'importe quelle occurrence du terme t_j et un autre terme t_k de la requête, avec $position^{e_i}(t_j)$ la position du terme t_j dans e_i. Une fontion finale $PC(t_j)$ est ensuite utilisée pour sélectionner et pondérer les termes pertinents :

$$PC(t_j) = P(t_j|E^r) \times \sum_{i=1}^{p_{e_j}} context^{e_i}(t_j) \quad (1.11)$$

où p_{e_j} est l'ensemble des éléments pertinents contenant le terme t_j.

Les termes ainsi sélectionnés et pondérés sont ajoutés finalement à la requête initiale (requête CO ou CAS), en normalisant leur poids entre 0 et 1.

On trouvera le détail de nos expérimentations dans (Hlaoua et al., 2007c,d,e, 2010). D'autres propositions pour la pondération des termes, notamment liées à la prise en compte de la pertinence négative, y sont présentées. Les résultats ont montré que :

- le nombre de termes réinjectés dans la requête initiale ne provoque pas une grande variance des résultats,
- le contexte des termes utilisé pour la sélection et la pondération des termes permet d'obtenir des améliorations optimales (comparé à Rocchio ou à la simple utilisation de l'équation 1.7),
- l'utilisation de la pertinence négative n'a pas d'intérêt.

Approche orientée structure L'idée principale derrière notre approche orientée structure est que les informations pertinentes recherchées par un utilisateur se retrouvent probablement dans des éléments de même type (c'est-à-dire possédant la même balise). Nous avons vérifié cette intuition sur les collections INEX 2005 et 2006 (Hlaoua and Pinel-Sauvagnat, 2006; Hlaoua et al., 2006b) et montré que 3 balises représentent à elles-seules plus de 90% des éléments pertinents.

Nous avons donc proposé l'algorithme SCA - *Smallest Common Ancestor* (Algorithme 2) (Hlaoua et al., 2006a). Les notations suivantes sont utilisées : E^r est l'ensemble des éléments pertinents, $e_i \in E^r$ est caractérisé par un chemin XPath simplifié S_i et un score w_i initialisé à 1 au début de l'algorithme. S_i est seulement composé de balises (exemple : `/destination/environment/fauna`). $S_i.first$ et $S_i.last$ sont respectivement la première et dernière balise de la structure S_i. $head(S)$ est une fonction permettant de réduire le chemin S_i en lui attribuant celui du parent (c'est-à-dire en supprimant la dernière balise de la structure). L'algorithme consiste à comparer la structure de chaque élément pertinent avec le reste des structures des éléments jugés pertinents. Pour chaque $(e_i, e_j) \in E^p \times E^p$, nous appliquons la fonction SCA qui permet d'extraire le chemin du plus petit ancêtre commun entre e_i et e_j. Le chemin est ensuite ajouté à un ensemble de structures communes CS. Pour exprimer la nouvelle requête, on sélectionne dans l'ensemble CS la ou les structures ayant le plus grand score. Ces structures seront utilisées sous une forme simplifiée qui correspond au $S_i.last$, et sont ajoutées à la requête initiale.

On trouvera le détail des expérimentations menées sur les collection INEX 2005 et 2006 dans (Hlaoua and Pinel-Sauvagnat, 2006; Hlaoua et al., 2006b, 2010). D'une manière générale,

```
SCA(e_i, e_j);
if S_i.first = S_j.first then
    if S_i.last = S_j.last then
        if ∃e_p(S_p, w_p) ∈ CS/S_p = S_i then
            w_p ← w_p + w_j ;
        else
            w_i ← w_i + w_j;
            CS ← (S_i, w_i) ;
        end
    else
        if head(S_j) ≠ null then
            S'_j ← head(S_j);
            w'_j ← w_j/2;
            SCA(e_i(S_i, w_i), e'_j(S'_j, w'_j));
        else
            SCA(e_j, e_i);
        end
    end
end
```

Algorithme 2 : SCA - *Smallest Common Ancestor*. Extraction de structures pertinentes

les résultats montrent l'intérêt de l'injection de nouvelles structures dans les requêtes, que ce soit des requêtes CO ou CO+S. Les meilleurs résultats sont obtenus par l'intégration de 3 structures, ce qui est cohérent avec nos observations préliminaires, à savoir qu'il existe des structures pertinentes. Ceci montre également que nous sommes capables de les identifier. De plus, les améliorations également obtenues sur les requêtes CO+S montrent que l'utilisateur n'est que rarement capable d'identifier seul les structures pertinentes dans sa requête.

Approche orientée contenu et structure Une suite logique de ces résultats est de combiner notre approche orientée contenu avec notre approche orientée structure. La nouvelle requête sera enrichie à la fois de conditions de structure et de contenu. Une combinaison naïve consiste à ajouter les termes extraits aux termes d'origine, tout en les encapsulant dans les structures identifiées comme pertinentes (Hlaoua et al., 2007a). De façon plus aboutie, nous avons proposé une combinaison flexible permettant de mettre en avant les relations contextuelles existant entre les termes pertinents et les structures pertinentes. L'idée est d'ajouter les termes dans les structures dans lesquelles ils apparaissent dans la collection. Pour ce faire, pour chaque terme pertinent, nous calculons la somme de ses occurrences dans les éléments ayant le même label l divisée par sa fréquence totale. Ce facteur est appelé partition $Part(t_i, l)$:

$$Part(t_i, l) = \frac{\sum_{j=1}^{N} Occ(t_i, e_j)}{|e_j|} \tag{1.12}$$

où N est le nombre d'éléments de label l dans lesquels on trouve t_i, et $Occ(t_i, e_j)$ est le nombre d'occurrences de t_i dans e_j.
Le poids de chaque terme extrait dans chaque structure extraite est finalement évalué selon la formule suivante :

$$score(t_i, l) = W_{t_i}/timesPart(t_i, l) \tag{1.13}$$

où W_{t_i} est évalué selon une des méthodes vue précédemment (équations 1.7 et 1.11). Les résultats présentés dans (Hlaoua et al., 2007b, 2010) montrent des performances encourageantes de la méthode, même si l'amélioration dépend de la collection considérée. Il reste

néanmoins clair qu'il existe une relation contextuelle entre les termes pertinents et les structures pertinentes.

Les travaux que nous venons de présenter s'appuient tous, de manière plus ou moins forte, sur l'information de structure contenue au sein des documents XML afin de retrouver des granules textuels. L'information textuelle n'est cependant pas la seule contenue dans les documents XML. On peut y trouver des contenus multimédia, dont la recherche a fait l'objet d'une autre partie de nos travaux.

1.2.2 Recherche de granules multimédia

Dans un contexte de recherche de contenus multimédia au sein de documents XML, deux types de résultats peuvent être renvoyés à l'utilisateur (Tsikrika and Westerveld, 2008; Westerveld and van Zwol, 2007a) :

— des éléments multimédia, c'est-à-dire des objets multimédia comme des images par exemple. Ces éléments multimédia réfèrent plus spécifiquement aux objets multimédia via les noms de fichiers associés, et ils peuvent également contenir d'autres éléments ayant des informations spécifiques sur le contenu (comme une légende par exemple).

— des fragments multimédia, composés d'objets multimédia et de texte associé. Ils peuvent être vus comme des granules documentaires contenant au moins un objet multimédia.

Si l'on considère le document XML de la figure 1.8 par exemple, le nœud `/destination/environment/fauna/image` est un élément multimédia. Ce dernier est composé de deux autres éléments contenant des informations spécifiques à l'image : `caption` et `illustrator`. Les nœuds `/destination/environment/fauna`, `destination/environment` ou encore `/destination` sont quant à eux considérés comme des fragments multimédia. Un fragment multimédia peut donc contenir plusieurs éléments multimédia. La figure 1.9 donne un aperçu d'une visualisation possible des deux types de granules pour l'utilisateur.

Dans ce contexte, nos recherches se sont donc focalisées sur deux axes :

— l'utilisation du contexte textuel et structurel des éléments multimédia pour évaluer leur pertinence,

— la détermination du meilleur fragment à renvoyer étant donné un élément pertinent (granularité appropriée).

Même si nos travaux peuvent être appliqués à tout type de média, nous nous sommes focalisés sur la recherche d'images, média le plus couramment inclus dans les documents XML.

La littérature sur la recherche de contenus multimédia peut être divisée en deux grandes classes. Les approches orientées contenu se basent sur un ensemble de caractéristiques de bas niveau comme la texture, la distribution des couleurs, les formes en recherche d'image, ou le ton et le timbre en recherche audio. Les approches orientées contexte utilisent le contexte des objets multimédia pour déterminer leur pertinence. Nous avons conduit des premiers travaux cherchant à combiner des systèmes orientés contenu avec des approches simples orientées contexte lorsque la requête était une image exemple ou une combinaison d'image exemple et de texte (Torjmen et al., 2008e, 2009a). Ces premiers résultats ont montré l'intérêt de combiner les deux types d'approches dans ce contexte de recherche. Nos travaux suivants se sont ensuite uniquement focalisés sur des requêtes orientées contenu ou contenu et structure. Dans la littérature, la plupart des approches estiment qu'il y a un lien fort entre les images et le texte les entourant dans les documents (Gong et al., 2006; Noah et al., 2008). Le contexte peut également être élargi à d'autres documents, en utilisant par exemple les hyperliens (Harmandas et al., 1997).

Les documents XML et leur information de structure inhérente, ont ouvert la voix à de nouveaux modèles, cherchant à rendre compte de l'importance du contexte textuel via les

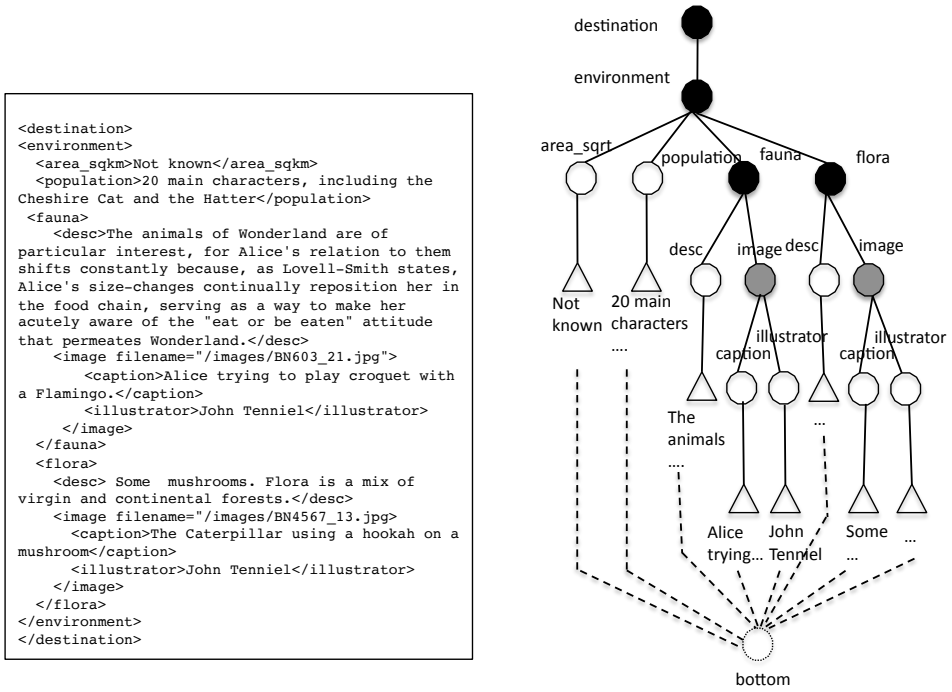


FIGURE 1.8 – Différence entre fragment et élément multimédia. Exemple sur l'arbre de la figure 1.1. Les nœuds grisés sont des éléments multimédia, les nœuds en noir sont des fragments.

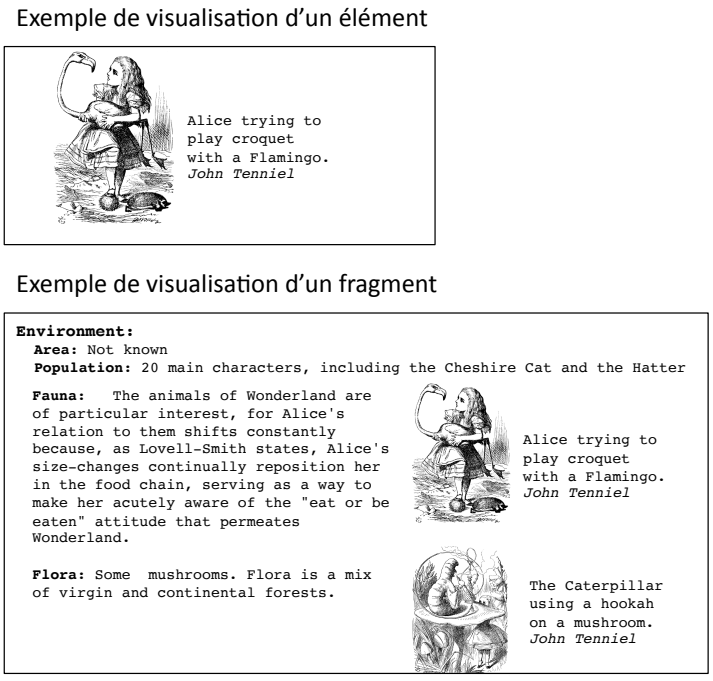


FIGURE 1.9 – Exemple de visualisation de fragment et élément multimédia

structures contenantes. Parmi les approches existantes de la littérature, certaines propositions utilisent des approches de RI structurée et filtrent ensuite simplement les résultats pour

Ontologie	Représentation arborescente du document
Concepts de l'ontologie	Nœuds du document XML
Relation « est partie de » entre les concepts	Relation hiérarchique entre les nœuds
Mesure de similarité entre 2 concepts C_1 et C_2 $Sim(C_1, C_2)$	Mesure de participation d'un nœud textuel ln dans la représentation de l'élément multimédia me $\phi(me, ln)$

TABLEAU 1.1 – Analogie entre une ontologie et un document XML

vérifier la présence de contenu multimédia (Kong and Lalmas, 2007b; Tsikrika et al., 2007). Une approche plus spécifique est celle de Kong and Lalmas (2007a) qui divise le document en régions de connaissance (*Region Knowledge*) autour de l'élément multimédia (ancêtre, frères, ...) afin d'établir sa pertinence. Cependant, même si la méthode exploite la structure verticale des documents, elle ne prend pas en compte la position des éléments contenus dans une même région (la structure horizontale n'est donc pas utilisée).

1.2.2.1 Recherche d'éléments multimédia

Afin d'utiliser au mieux l'information de structure au sein des documents XML, nous avons proposé deux approches principales pour la recherche d'éléments multimédia :

- l'approche **CBA** se base sur les nœuds proches des éléments multimédia afin d'évaluer leur pertinence (Hlaoua et al., 2007e; Torjmen et al., 2007b, 2009c) : les nœuds fils (**C**hildren), frères (**B**rothers) et les nœuds ancêtres (**A**ncestors) ayant un score de pertinence (d'après le modèle XFIRM) non nuls sont utilisés pour évaluer la pertinence de chaque élément multimédia.
- l'approche *OntologyLike* se base sur l'idée que chaque nœud textuel du document possède une relation sémantique avec l'élément multimédia, et fait l'analogie entre l'arbre d'un document XML et une ontologie.

L'approche CBA, bien qu'obtenant des résultats très satisfaisants sur les campagnes d'évaluation INEX 2006 et INEX 2007 (*Multimedia Task*) est fortement dépendante du système de recherche d'information utilisé au départ, ainsi que de nombreux paramètres (Torjmen et al., 2010). L'approche **OntologyLike** pallie ces inconvénients, en faisant l'analogie entre arbre du document et ontologie. Les nœuds de l'arbre peuvent être considérés comme des concepts, et la structure hiérarchique entre ces nœuds peut être considérée comme la relation hiérarchique « est partie de » (un élément `paragraph` est partie de `section` par exemple). En suivant cette analogie, nous nous sommes inspirés des mesures de similarité sémantique entre les concepts pour déterminer une mesure rendant compte du degré de participation de chaque nœud textuel dans l'évaluation de la pertinence de l'élément multimédia (Torjmen and Pinel-Sauvagnat, 2009; Torjmen et al., 2008a,b). Le tableau 1.1 résume notre analogie.

La pertinence d'un élément multimédia me par rapport à une requête q est évaluée de la façon suivante :

$$S(me, q) = \sum_{ln_i \in L_{doc(me)}} \phi(me, ln_i) \times RSV(ln_i, q) \tag{1.14}$$

Le facteur $RSV(ln_i, q)$ décrit la pertinence textuelle des nœuds feuille ln_i (évaluée dans notre cas avec le modèle XFIRM) et le facteur ϕ reflète la proximité des nœuds feuille à l'élément multimédia considéré. Intuitivement, nous désirons que le facteur ϕ reflète les faits suivants : les éléments fils de l'élément multimédia doivent participer plus à sa pertinence

que ses ancêtres et ses frères, et ces derniers doivent participer plus que les descendants textuels des ancêtres. En effet, les éléments textuels descendants peuvent être considérés comme les nœuds les plus spécifiques pour représenter l'élément multimédia, les éléments textuels descendants des frères ont une grande probabilité de partager le même sujet et les autres nœuds permettent de prendre en compte le contexte du document dans son entier.

Nous avons évalué plusieurs mesures de similarité sémantique de la littérature basées sur le nombre d'arc pour traduire le facteur ϕ, parmi lesquelles Rada (Rada et al., 1989) ou encore Wu-Palmer (Wu and Palmer, 1994). Même si ces mesures permettent de différencier l'importance des nœuds textuels, certains nœuds frères pourraient avoir plus d'influence dans le score de pertinence final que les nœuds descendants, ce qui va à l'encontre de notre intuition. Nous avons donc cherché à traduire notre intuition par une nouvelle mesure de similarité (Torjmen and Pinel-Sauvagnat, 2009; Torjmen et al., 2008a,b, 2010). Nous avons retenu la formule suivante, inspirée de (Zargayouna, 2004) :

$$\phi_{OntLike}(me, ln_i) = \frac{1}{(N_1 + w) * N_2 * depth(CS(me, ln_i))} \tag{1.15}$$

où $N_1 = dist(me, CS)$ et $N_2 = dist(ln_i, CS)$ sont les distances séparant me et ln_i de leur plus proche ancêtre commun CS. $depth(CS(me, ln_i))$ est le nombre maximum d'arcs entre $CS(me, ln_i)$ et le nœud *bottom*, qui un concept virtuel reliant tous les nœuds feuille. Ce nœud *bottom* est représenté sur la figure 1.8. w ($w > 0$) est ajouté à N_1 pour éviter la division par zéro lorsque l'élément multimédia lui-même est le plus proche ancêtre commun entre le nœud textuel et l'élément multimédia.

1.2.2.2 Recherche de fragments multimédia

Lorsque la recherche porte sur des fragments multimédia, nous évaluons le score de pertinence de chaque fragment multimédia de la façon suivante :

$$\begin{aligned} S(mf, q) &= \lambda * S_{XFIRM}(mf, q) \\ &\quad +(1 - \lambda) * \sum_{i=1}^{|me|} \theta * S(me_i, q) \end{aligned} \tag{1.16}$$

où $S_{XFIRM}(mf, q)$ est le score de contenu du fragment multimédia, évalué avec le modèle XFIRM, et θ prend en compte la distance entre les éléments multimédia me_i et le fragment multimédia mf :

$$\theta = \frac{1}{Dist(me_i, mf) + 1} \tag{1.17}$$

ou

$$\theta = K^{(Dist(me_i, mf)+1)} \tag{1.18}$$

1.2.2.3 Discussion

Nos approches, tant pour la recherche d'éléments multimédia que de fragments multimédia, ont été évaluées grâce aux campagnes INEX 2006 et 2007 (tâche *Multimedia*) (voir (Torjmen et al., 2009c) et (Torjmen-Khemakhem et al., 2013) pour une synthèse de ces expérimentations). D'une façon générale, nous avons montré que la prise en compte de la structure aide à l'amélioration des performances :

— dans le cas de la recherche d'éléments multimédia, la structure est utilisée pour pondérer l'importance des différents granules textuels dans la pertinence des éléments. Ceci permet de donner plus ou moins d'importance à l'information textuelle en fonction de sa proximité structurelle avec l'élément.

— dans le cas de la recherche de fragments, la structure est utilisée via un facteur de distance θ entre le fragment multimédia et les éléments multimédia associés. Là encore, nous avons montré que les éléments de contexte d'un fragment ne doivent pas tous être considérés de la même façon.

Si l'on considère maintenant la performance de nos approches par rapport à l'état de l'art, nous nous classons systématiquement dans le Top 5 en nous comparant avec les participations officielles d'INEX, et même à la première place sur certaines mesures d'évaluation (iP[0.01] par exemple, voir (Torjmen-Khemakhem et al., 2013)). En 2007, les requêtes de la tâche *Multimedia* ont été intégrées à l'ensemble des requêtes de la tâche *Adhoc*. Les *runs* officiels de la tâche *Adhoc* ont donc été évalués dans le contexte de la tâche *Multimedia*, et de manière surprenante, ont obtenus de meilleurs résultats. Nos résultats actuels permettent cependant d'atténuer ce constat, puisque de meilleures performances sont observées avec notre système.

1.3 Synthèse des travaux présentés

Nos travaux dans le cadre de la RI structurée se sont donc principalement attachés à démontrer l'intérêt de l'information de structure dans la représentation de la pertinence :

— d'un point de vue requête, nous avons montré l'intérêt d'ajouter des conditions de structure aux requêtes qui n'en possèdent pas, montrant ainsi que l'utilisateur n'est pas forcément le meilleur pour exprimer son besoin. Ces travaux sont à notre connaissance parmi les premiers du domaine pour la reformulation de requêtes dans des documents XML (Hlaoua et al., 2007b, 2010).

— d'un point de vue modèle,

  — nous avons montré qu'il existe une relation contextuelle entre les termes pertinents et les structures pertinentes (Hlaoua et al., 2006b),

  — nous avons proposé des modèles pour la recherche de granules textuels ou multimédia, prenant tous en compte la structure des documents. La structure est ainsi considérée soit de façon implicite (méthode de propagation et rétro-propagation de la pertinence pour la recherche de granules textuels, méthode *CBA* pour les granules multimédia (Torjmen et al., 2009c)), soit de façon explicite (distance d'édition et modèles de langues pour les granules textuels en réponse à des requêtes très structurées (Laitang et al., 2013a,b), méthode *OntologyLike* pour les granules multimédia (Torjmen-Khemakhem et al., 2013)).

Nos approches ont été évaluées de façon systématique sur des collections de test de référence (voir l'annexe A et les tableaux A.1 et A.2), ainsi que lors de participations officielles aux campagnes d'évaluation du domaine :

— INEX 2005 (Sauvagnat et al., 2005), 2006 (Hlaoua et al., 2007e), 2007 (Torjmen et al., 2007a), 2011 (Laitang et al., 2012a),

— CLEF 2007 (Torjmen et al., 2008e), 2008 (Torjmen et al., 2009a,b).

Si l'on compare nos résultats avec l'état de l'art du domaine, nos approches pour le traitement des requêtes très structurées (distance d'édition et modèles de langue basés sur l'information structurelle) nous permettent de nous classer au-dessus des premiers participants aux campagnes d'évaluation INEX 2005 *SSCAS* et INEX 2010 *Datacentric* sur les principales mesures officielles. Les résultats de nos approches pour la recherche multimédia nous

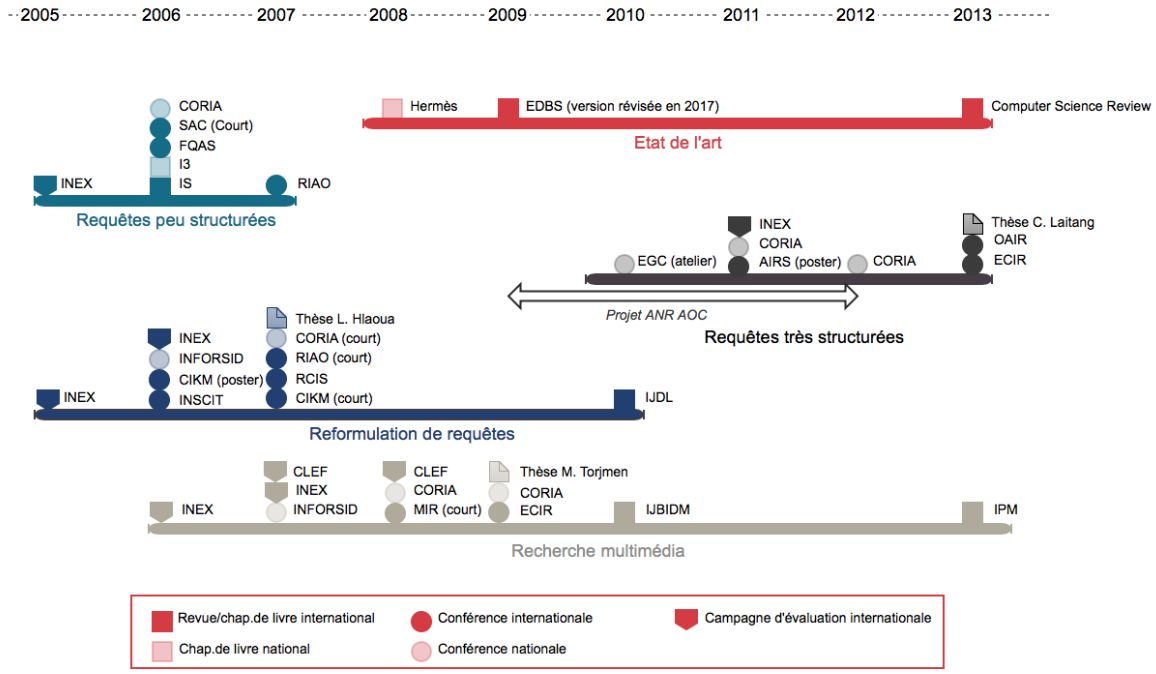


FIGURE 1.10 – Graphique synthétique de la structuration et de la valorisation des travaux sur le thème Recherche de granules XML.

permettent quant à eux de nous classer systématiquement dans le Top 5 des tâches INEX *Multimedia* 2005 et 2006, et même à la première place sur certaines mesures d'évaluation [5].

Diffusion scientifique. La figure 2.4 présente de façon chronologique mes différentes publications citées dans le chapitre, organisées en fonction des problématiques abordées.

Formation à la recherche. J'ai co-encadré ou encadré trois étudiants de Master sur ces thématiques (Boujot, 2007; Le, 2009; Torjmen, 2006).

J'ai également participé à l'encadrement, avec Mohand BOUGHANEM, de trois thèses, dont les contributions ont été largement décrites dans le chapitre :

— Lobna HLAOUA (Hlaoua, 2007),

— Mouna TORJMEN (Torjmen, 2009),

— Cyril LAITANG (Laitang, 2013).

Projets. Une partie de ces recherches a été menée dans le cadre du projet ANR AOC (*Appariement d'Objets Complexes* [6]), qui s'est déroulé de 2009 à 2012. Nous nous sommes principalement intéressés dans ce projet à l'appariement d'arbres XML en utilisant des algorithmes d'appariement d'arbres issus de la théorie des graphes.

5. Ces comparaisons à l'état de l'art sont faites a posteriori, en réutilisant les collections de test fournies par les campagnes d'évaluation.

6. https://aoc.irit.fr/, dernier accès en février 2018

CHAPITRE 2

Recherche de granules d'information : microblogs et information temps-réel

Oh dear! Oh dear! I shall be late!

— The White Rabbit

"If you knew Time as well as I do," said the Hatter, "you wouldn't talk about wasting it. It's him."

Sommaire

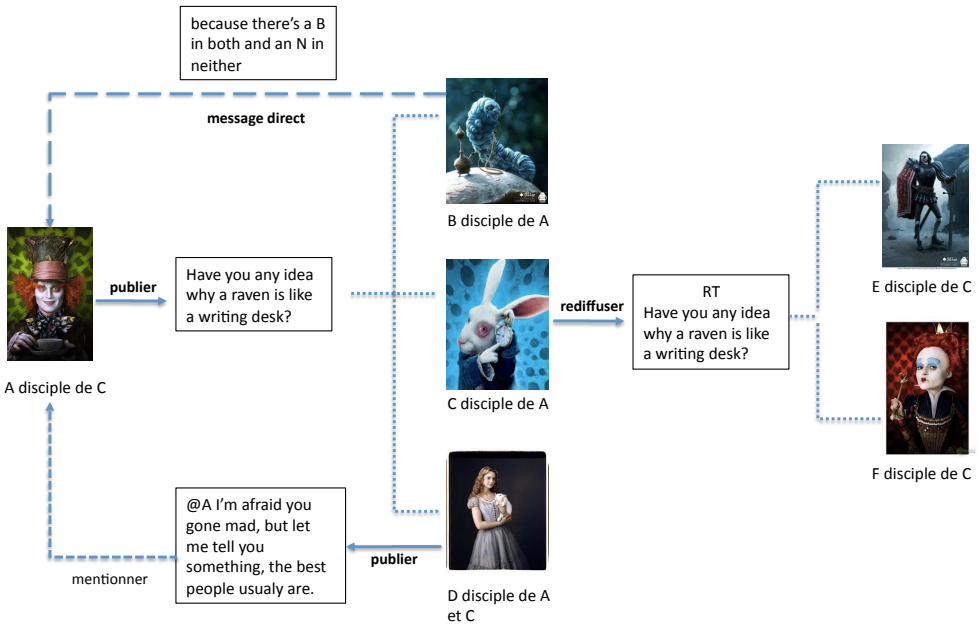


FIGURE 2.1 – Fonctionnement de la plateforme de microblogging Twitter

2.1 État de l'art et problématiques ciblées

2.1.1 Introduction

Les plateformes de *microblogging* telles que Twitter [1], Sina Weibo pour la Chine [2] ou encore Tumblr [3] sont devenues en quelques années une source d'information incontournable pour les internautes. En témoignent les 2,1 milliards de requêtes soumises tous les jours au moteur de recherche de la plateforme Twitter [4], ou encore l'annonce début 2015 de Google indiquant la ré-intégration du flux Twitter dans ses résultats de recherche [5].

La quantité gigantesque d'information publiée tous les jours par les microblogueurs (700 millions de tweets par jour par exemple début 2018 [6]), rend cruciale la problématique de l'accès à l'information. Cette problématique a été prise sérieusement en compte par la communauté de la RI, comme le montrent les (très) nombreuses publications scientifiques sur le sujet.

Le schéma 2.1, adapté de (Damak, 2014), synthétise le fonctionnement de la plateforme de microblogging Twitter, qui sert de cas d'application à nos travaux.

Un microblogueur (un individu, une société ou encore un site d'information) peut s'abonner au flux d'autres microblogueurs (il devient un *follower*, c'est-à-dire un disciple de ces utilisateurs). Cet abonnement ne nécessite pas l'autorisation des utilisateurs concernés, et lui permet de suivre sur sa page d'accueil (sa *timeline*) toutes les publications (les *tweets*) des personnes qu'il suit. Cette timeline affiche les messages par ordre chronologique inverse de leur arrivée, c'est-à-dire que les plus récents sont affichés en premier. Lorsqu'un utilisateur publie

1. http://www.twitter.com, dernier accès mars 2018
2. http://www.weibo.com, derniers accès mars 2018
3. https://www.tumblr.com/, dernier accès mars 2018
4. Source : http://www.statisticbrain.com/twitter-statistics/, accédé le 12/02/2018.
5. Source :http://searchengineland.com/report-google-will-get-access-twitters-firehose-214220, accédé le 12/02/2018.
6. Source : http://www.internetlivestats.com/, accédé le 12/02/2018.

FIGURE 2.2 – Exemple de tweet

un tweet, tous ses followers le voient donc, sauf si ce tweet est un message privé. Lorsqu'un tweet est rediffusé, on parle de *retweet*, et le message porte la mention RT. Un tweet peut également mentionner/s'adresser à des utilisateurs particuliers, qui sont alors directement cités dans le tweet (*@mention*).

L'accès à l'information sur les plateformes de microblogging peut se faire de différentes façons (Damak, 2014) :

— par la recherche des microblogs (Ounis et al., 2011),

— par la recherche de microblogueurs (Bartoletti et al., 2016; Ben Jabeur et al., 2012b; Weng et al., 2010; Yamaguchi et al., 2010),

— en faisant de la détection d'opinions (Dubois and Gaffney, 2014; Khan et al., 2014; Pak and Paroubek, 2010),

— ou encore en faisant de la détection de tendances (Zubiaga et al., 2015) ou d'événements (Atefeh and Khreich, 2015).

Si l'on s'intéresse plus particulièrement à la recherche de microblogs, Teevan et al. (2011) ont montré que les plateformes de microblogging sont avant tout utilisées pour retrouver de l'information concise et temps-réel sur des événements récents. La compréhension des caractéristiques spécifiques des tweets est fondamentale pour toute approche s'intéressant à leur recherche. La figure 2.2 montre un exemple de tweet.

Ce tweet fait mention via le signe @ des utilisateurs `Alice`, `theHatter`, et `theWhiteRabbit`. Son texte est très court (il ne doit pas dépasser 140 caractères [7]). Il est souvent (ce n'est pas le cas ici) exprimé dans un langage spécifique, à la mode SMS. Le tweet contient également deux *hashtags*, dénotés par le signe #. Un hashtag indique un mot important pour l'auteur du tweet, qui peut ensuite servir lors d'une recherche directe dans la plateforme. Le tweet contient également une URL. Les liens sont souvent donnés avec une forme courte, générée par des services tels que `bit.ly` ou `tinyurl.com`, en raison du nombre limité de caractères autorisés dans un tweet. Lorsque l'URL fait référence à un tweet ou encore une image (c'est le cas ici), le contenu s'affiche directement sous le tweet.

Outre le contenu du tweet, des méta-données sont associées à chaque publication :

— l'auteur du tweet (ici `CheshireCat`), auquel est associé un profil consultable par les utilisateurs de la plateforme,

— ses jours et heures de publication,

7. Cette limite a été étendue à 280 caractères fin 2017.

— le nombre de fois où le tweet a été aimé ou retweeté (respectivement 111 fois et 15 fois dans notre exemple),

— la géolocalisation associée, si l'auteur l'a activée dans le cas d'une publication sur téléphone mobile ou tablette,

— l'information éventuelle du fait que le tweet est une rediffusion (un *retweet*).

D'un point de vue RI « pure », les spécificités remarquables sont donc les suivantes :

— une taille réduite à quelques termes et donc un vocabulaire limité,

— une qualité de langage relativement pauvre et différente des publications traditionnelles (langage parlé/SMS, abréviations,...),

— une syntaxe spécifique (@mention et #hashtag),

— une publication temps-réel et un flux continu d'information.

2.1.2 Problématiques ciblées

Une première analyse de défaillance des modèles de recherche « traditionnels » effectuée sur les collections TREC Microblog 2011 et 2012 nous a permis de confirmer/mettre en évidence les limites suivantes, principalement liées à des problèmes de vocabulaire (Damak, 2014) :

— la concision des microblogs engendre une correspondance limitée entre les termes des requêtes et ceux des microblogs (absence totale des termes de certaines requêtes dans les microblogs pertinents, noms propres et entités nommées orthographiés de différentes façons...) ;

— une lemmatisation pas ou peu efficace, due au langage utilisé et aux termes concaténés pour former des #hashtags ;

— la non prise en compte du caractère temps-réel de la recherche.

De façon quasi-systématique, l'appariement entre une requête composée de 3 ou 4 termes et un microblog composé d'en moyenne 15 termes revient à vérifier la présence ou l'absence des termes de la requête dans le microblog.

Nos travaux se sont donc focalisés sur les problématiques suivantes :

— comment pallier la longueur limitée des tweets au niveau des modèles de recherche ?

— quels sont les facteurs (sources d'évidence) autres que le contenu qui reflètent la pertinence ?

— comment prendre en compte la fraîcheur des tweets dans la recherche ?

2.2 Contribution au domaine de recherche

Les propositions que nous présentons ici sont principalement empiriques. On trouvera en Annexe 1, Tableau A.3, la description des collections de test utilisées.

2.2.1 Expansion de requêtes et documents

Afin de résoudre le problème de non-correspondance de vocabulaire, nous avons mis en œuvre plusieurs méthodes d'expansion de requêtes et de documents (Ben Jabeur et al., 2013; Damak, 2014).

2.2.1.1 Expansion de requêtes

Plusieurs possibilités ont été évaluées :

— l'expansion des requêtes via des articles d'actualité : l'idée a été de construire un mega-document constitué de la concaténation des 5 premiers articles du NYTimes et du Gardian renvoyés en réponse à la requête et publiés avant la date de cette dernière, et d'en extraire via l'API Alchemy [8] les mots-clés les plus importants, ajoutés ensuite à la requête,

— l'expansion des requêtes avec le premier *synset* retrouvé par la base lexicale WordNet pour chaque terme de la requête,

— l'expansion des requêtes via les suggestions orthographiques du moteur de recherche Bing,

— l'expansion des requêtes par réinjection de pertinence, en utilisant une version modifiée (Salton and Buckley, 1997) de la formule de Rocchio (Rocchio, 1971) ou bien en utilisant le mécanisme naturel de la formule du BM25.

2.2.1.2 Expansion des microblogs

Cette expansion a été faite :

— en rajoutant à chaque tweet contenant un hashtag les termes concaténés formant le hashtag (par exemple `#ParExemple` permet d'ajouter les termes `par` et `exemple`) ;

— en rajoutant à chaque tweet contenant une URL le contenu de l'URL.

2.2.1.3 Résultats

Le protocole complet de test ainsi que nos résultats détaillés sur les collections TREC 2012 et 2013 se trouvent dans (Ben Jabeur et al., 2013; Damak, 2014). De manière résumée, les expérimentations ont montré que l'extension des tweets avec le contenu de leurs URLs et l'expansion de requêtes via Rocchio permettait d'augmenter le rappel. L'utilisation de sources externes (articles d'actualité, Wordnet, Bing) n'a quant à elle pas montré son intérêt.

2.2.2 Facteurs de pertinence pour la RI dans des microblogs

L'état de l'art sur la recherche adhoc de tweets montre que les spécificités des tweets (taille, vocabulaire et syntaxe spécifiques) ont été prises en compte via la combinaison de facteurs divers venant s'ajouter à la pertinence du contenu. Le tableau 2.1, tiré de (Damak et al., 2013) présente ces facteurs, en indiquant pour chacun si des ressources externes ou des informations du futur sont utilisées pour les évaluer.

Nous avons souhaité comprendre l'impact de ces facteurs, et avons pour cela effectué la sélection de 14 facteurs présentés dans les paragraphes ci-dessous [9]. Soit q une requête composée de mots-clés (*topic*) associée à une date (*timestamp*) tmp_q, C_q le corpus de tweets publiés avant d_q, T_q l'ensemble des tweets restitués par un moteur de recherche donné utilisant uniquement le contenu des tweets pour évaluer leur pertinence ($T_q \in C_q$), et t_i un tweet $\in C_q$ sur lequel on applique le facteur de pertinence.

2.2.2.1 Facteurs de pertinence étudiés

1. Facteurs basés sur le contenu des tweets

8. http://www.alchemyapi.com, jusqu'en 2015, maintenant https://www.ibm.com/watson/services/natural-language-understanding/, dernier accès mars 2018.

9. On notera que certains facteurs ont été supprimés ou adaptés afin de ne pas tenir compte des évidences futures, non connues à l'instant de la requête.

Facteur		E	F
Popularité du tweet	(Duan et al., 2010)	-	-
Nombre de termes de la requête dans le tweet	(Damak et al., 2011)	-	-
Nombre de re-tweets	(Duan et al., 2010; Magnani et al., 2012; Vosecky et al., 2012; Zhao et al., 2011)	-	+
Fréquence des hashtags	(Duan et al., 2010)	-	-
Présence de hashtag	(Metzler and Cai, 2011; Vosecky et al., 2012)	-	-
Popularité du hashtag	(Vosecky et al., 2012)	-	+
Longueur du tweet	(Duan et al., 2010; Magnani et al., 2012; Metzler and Cai, 2011; Zhao et al., 2011)	-	-
Présence d'URL	(Duan et al., 2010; Massoudi et al., 2011; Metzler and Cai, 2011; Vosecky et al., 2012)	-	-
Nombre d'URLs dans le tweet	(Zhao et al., 2011)	-	-
Popularité de l'URL	(Vosecky et al., 2012)	-	+
Est une réponse	(Duan et al., 2010; Metzler and Cai, 2011; Vosecky et al., 2012)	-	-
Nombre de tweets de l'utilisateur	(Zhao et al., 2011)	-	+
Nombre de followers	(Duan et al., 2010; Magnani et al., 2012; Massoudi et al., 2011; Zhao et al., 2011)	-	+
Nombre de mentions	(Duan et al., 2010; Vosecky et al., 2012)	-	+
Fraîcheur	(Magnani et al., 2012; Metzler and Cai, 2011; Vosecky et al., 2012)	-	-
Qualité du langage	(Metzler and Cai, 2011)	+	-

TABLEAU 2.1 – Facteurs utilisés dans la littérature pour déterminer la pertinence des tweets. (E) dénote un facteur externe et (F) un facteur utilisant une évidence du futur. Extrait de (Damak et al., 2013)

— Popularité du tweet (Duan et al., 2010) dans T_q : un tweet est considéré populaire si d'autres tweets ont un contenu similaire. La similarité $sim(t_i, t_j)$ entre chaque paire de tweets est calculée avec le modèle vectoriel (Cohen et al., 2007).

$$f_1(t_i, q) = \frac{\sum_{t_j \in T_q, i \neq j} sim(t_i, t_j)}{|T_q| - 1} \tag{2.1}$$

— Longueur du tweet (Duan et al., 2010) : On note $l(t_i)$ le nombre de termes de t_i.

$$f_2(t_i) = \frac{l(t_i)}{max_{t_j \in T_q} l(t_j)} \tag{2.2}$$

— Correspondance exacte des termes : ce facteur favorise les tweets qui contiennent les termes de la requête q. La valeur $nb(t_i, q)$ correspond au nombre de termes en commun entre t_i et q :

$$f_3(t_i, q) = \frac{nb(t_i, q)}{max_{t_j \in T_q} nb(t_j, q)} \tag{2.3}$$

— Qualité du langage (Duan et al., 2010) : ce facteur de pertinence représente la proportion des termes qui existent dans un dictionnaire [10] par rapport à tous les termes de t_i. La valeur $dic(term)$ est binaire : 1 si le terme existe dans le dictionnaire, 0 sinon :

$$f_{14}(t_i) = \frac{\sum_{term \in t_i} dic(term)}{l(t_i)} \tag{2.4}$$

2. Facteurs de pertinence basés sur l'hypertextualité

— Présence d'une URL dans le tweet (Nagmoti et al., 2010; Zhao et al., 2011) :

$$f_4(t_i) = \begin{cases} 1 & \text{si } t_i \text{ contient une URL} \\ 0 & \text{sinon} \end{cases} \tag{2.5}$$

— Nombre d'URLs dans le tweet (Zhao et al., 2011) :

$$f_5(t_i, q) = \frac{|\{w \in t_i / isURL(w)\}|}{max_{t_j \in T_q} |\{w \in t_j / isURL(w)\}|} \tag{2.6}$$

— Popularité de l'URL dans C_q

$$f_6(t_i, q) = \frac{\sum_{url \in t_i} freq(url)}{max_{t_j \in C_q} \sum_{url \in t_j} freq(url)} \tag{2.7}$$

3. Facteurs de pertinence basés sur les hashtags

— Présence de hashtag (Metzler and Cai, 2011).

$$f_7(t_i) = \begin{cases} 1 & \text{si } t_i \text{ contient un hashtag} \\ 0 & \text{sinon} \end{cases} \tag{2.8}$$

— Popularité du hahstag dans C_q (Duan et al., 2010). On note la fréquence d'un hashtag dans le corpus C_q par $freq(h)$:

$$f_8(t_i) = \sum_{h \in t_i} freq(h) \tag{2.9}$$

— Hashtags de la requête dans le tweet :

$$f_9(t_i, q) = \frac{|\{w \in q / \#w \in t_i\}|}{max_{t_j \in T_q} |\{w \in q / \#w' \in t_j\}|} \tag{2.10}$$

4. Facteurs de pertinence basés sur la popularité des auteurs

— Nombre de tweets de l'auteur (Nagmoti et al., 2010). On note par $a(t_i)$ l'auteur du tweet t_i et $N(a(t_i))$ le nombre de tweets publiés par l'auteur du tweet t_i dans le corpus C_q.

$$f_{10}(t_i) = N(a(t_i)) \tag{2.11}$$

— Nombre de citations de l'auteur (Zhao et al., 2011). Plus un auteur est mentionné, plus il est populaire. $M(a(t_i))$ indique combien de fois un auteur du tweet t_i a été mentionné dans le corpus C_q :

$$f_{11}(t_i) = M(a(t_i)) \tag{2.12}$$

5. Facteurs de pertinence relatifs à la qualité des tweets

10. http://code.google.com/p/language-detection/, dernier accès mars 2018.

— Retweet (Metzler and Cai, 2011).

$$f_{12}(t_i) = \begin{cases} 1 & \text{si } t_i \text{ contient RT} \\ 0 & \text{sinon} \end{cases} \tag{2.13}$$

— Fraîcheur (Magnani et al., 2012). C'est la différence entre la date de la publication du tweet t_i et la date de la soumission de la requête q, mesurée en secondes. $tmp(t_i)$ est le timestamp en seconde d'un tweet t_i (c'est-à-dire sa date de publication).

$$f_{13}(t_i, q) = \frac{tmp(q) - tmp(t_i)}{max_{t_j \in T_q} tmp(q) - tmp(t_j)} \tag{2.14}$$

Pour évaluer l'intérêt de ces facteurs, nous avons mené une première étude préliminaire en combinant leurs scores linéairement (Damak et al., 2012). Cette étude a mis en avant l'intérêt du facteur f_4 portant sur la présence d'URL dans le tweet. Afin d'approfondir les résultats, nous avons mené deux autres études (Damak et al., 2013) : une première basée sur les distributions de score des facteurs, et une seconde basée sur des approches de sélection d'attributs.

2.2.2.2 Étude basée sur la distribution des facteurs

La figure 2.3 montre les résultats obtenus en considérant les requêtes contenant plus de 100 tweets pertinents de la collection TREC 2011 (9 *topics* sur 49). Pour chacune de ces requêtes, nous avons pris tous les tweets pertinents auxquels nous avons ajouté le même nombre de tweets non pertinents. Les intervalles pour chaque facteur ont été calculés selon la règle de Sturges (Sturges, 1926).

On observe une distribution des scores différente entre les tweets pertinents et les tweets non pertinents pour les facteurs f1 (popularité du tweet), f2 (longueur du tweet), f3 (correspondance exacte des termes de la requêtes), f4 (présence d'URL), f5 (fréquences d'URLs), f6 (importance des URLS dans le corpus) et f13 (fraîcheur). Les tweets pertinents obtiennent de meilleurs scores sur ces facteurs, indiquant probablement que ces derniers reflètent la pertinence.

Approches par sélection d'attributs Afin de confirmer ces observations, nous avons utilisé des méthodes de sélection de facteurs (Hall and Holmes, 2003), dont le but est d'identifier et supprimer les facteurs non redondants et pertinents. Nous nous sommes appuyés sur l'outil d'apprentissage Weka [11] en suivant la méthodologie suivante : nous avons fait l'union des 1500 premiers tweets renvoyés par le moteur de recherche Lucene sur chacune des requêtes de la collection TREC 2011, obtenant ainsi 2129 tweets pertinents et 70614 tweets non pertinents d'après les *qrels* [12] associés à la collection. Ce résultat non équilibré sur la distribution des classes pourrait conduire les classifieurs à prédire les échantillons de la classe majoritaire en ignorant la classe minoritaire (Yen and Lee, 2006). Nous avons donc procédé à un sous-échantillonage pour réduire la classe des tweets non pertinents à 2129 tweets, en sélectionnant les tweets au hasard. Les résultats sont présentés dans le tableau 2.2.

Les facteurs mis en évidence par cette étude sont similaires aux précédents (f1, f2, f3, f4, f5, f6 et f13), ce qui confirme leur intérêt. Nous avons par la suite utilisé les facteurs sélectionnés par certaines des approches avec des algorithmes d'apprentissage associés (Hall and Holmes, 2003). Les résultats détaillés, disponibles dans (Damak et al., 2013), montrent que la sélection des facteurs en entrée des algorithmes d'apprentissage influe bénéfiquement sur les résultats. Si l'utilisation de techniques d'apprentissage basées sur ces facteurs permet d'obtenir des résultats satisfaisants par rapport aux méthodes de l'état de l'art (Ben Jabeur

11. http://www.cs.waikato.ac.nz/ml, dernier accès mars 2018.
12. Les *qrels* correspondent à la vérité terrain, ce terme est très utilisé dans le jargon TREC.

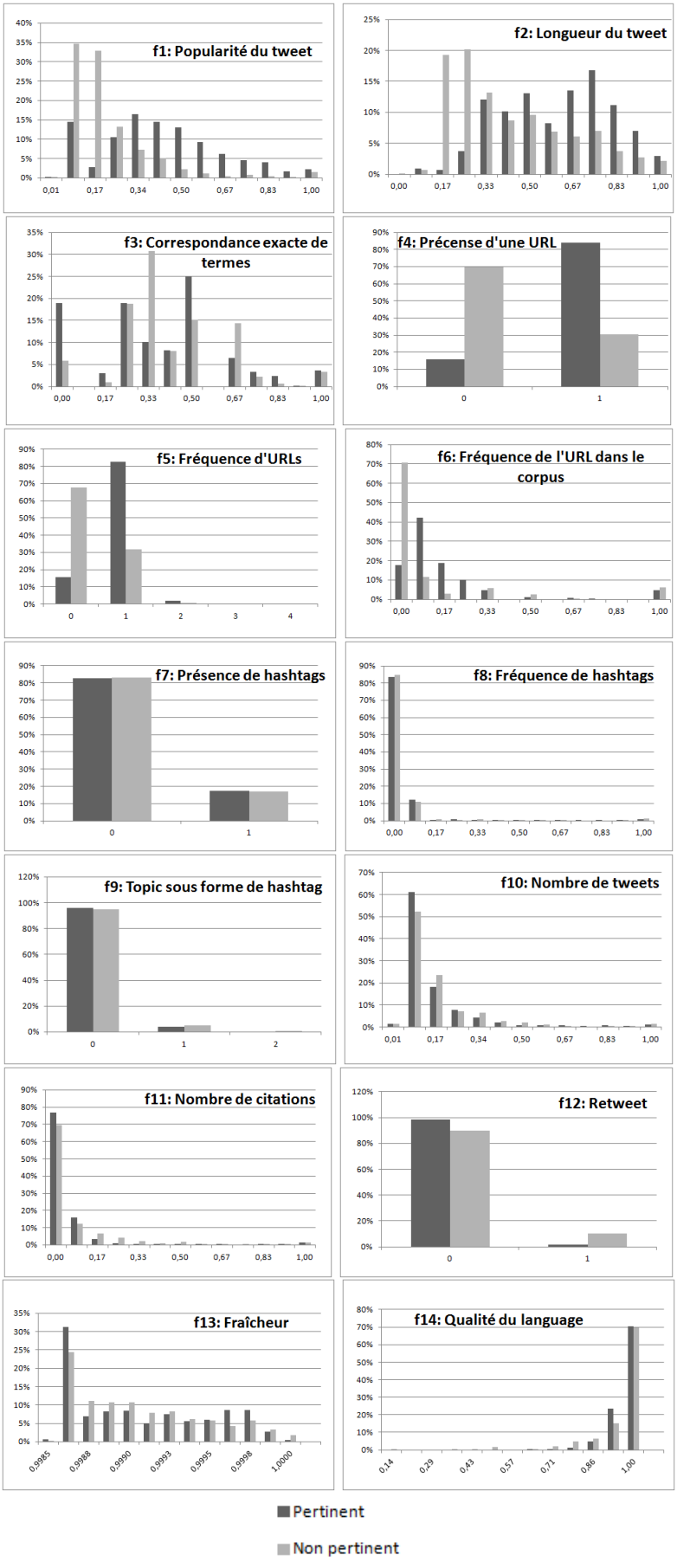


FIGURE 2.3 – Distribution des scores des tweets pertinents et des tweets non pertinents sur la collection TREC Microblog 2011. Extrait de (Damak, 2014).

Algorithme	f1	f2	f3	f4	f5	f6	f7	f8	f9	f10	f11	f12	f13	f14
Cfssubseteval	+	+	+	+	+	+						+	+	
ChisquaredAtt.Eval	+	+	+	+	+	+			+	+	+	+	+	+
FilteredAtt.Eval	+	+	+	+	+	+			+	+	+	+	+	+
FilteredSubsetEval	+	+	+	+	+	+							+	
Gain ration att eval	+	+	+	+	+	+			+	+	+	+	+	+
Info gain att eval	+	+	+	+	+	+			+	+	+	+	+	+
One att eval	+	+	+	+	+	+			+	+	+	+	+	+
ReliefFAttribute Eval	+	+	+	+	+	+			+	+	+	+	+	+
SVM Attribute Eval	+	+	+	+	+	+			+		+	+	+	+
SymetricalUncertEval	+	+	+	+	+	+			+	+	+	+	+	+
Consistency subset Eval	+	+	+	+	+	+			+	+	+	+	+	+
Wrapper subset Eval			+	+	+	+								
LatentSymanticAnalysis	+	+	+											
Total	12	12	13	12	12	12	0	0	9	8	9	10	11	9

TABLEAU 2.2 – Facteurs sélectionnés par des techniques de sélection d'attributs, collection TREC 2011 (Damak et al., 2013).

et al., 2012a), une simple combinaison linéaire des facteurs liés au contenu des tweets ainsi que ceux liés à l'hypertextualité avec le moteur de recherche Lucene nous permet d'obtenir les meilleurs résultats (Damak, 2014).

2.2.2.3 Discussion

Pour résumer, ces expérimentations nous ont permis de mettre en évidence l'intérêt prépondérant des facteurs relatifs aux URLs et au contenu du tweet. La longueur du tweet est notamment gage de son informativité. Les URLs, permettant d'ajouter du contenu aux tweets, permettent également de les discriminer. Les facteurs relatifs aux hashtags et à l'auteur du tweet semblent quant à eux sans intérêt pour la tâche de recherche adhoc. Le facteur basé sur la fraîcheur des tweets (f13) ne semble pas non plus prépondérant (il a pourtant a été utilisé 11 fois sur 13), de façon plus surprenante. La qualité temps-réel des tweets est pourtant une caractéristique clé, que nous discutons dans la section suivante.

2.2.3 Prise en compte du temps/de la fraîcheur

Comme nous l'avons dit en introduction de ce chapitre, la recherche temps-réel des tweets est une des motivations principales des utilisateurs de Twitter (Teevan et al., 2011). Nous avons ainsi mené d'autres expérimentations afin d'intégrer directement le facteur temps dans le modèle de recherche, en favorisant les tweets ou les termes récents, ou encore en prenant en compte la fréquence temporelle des termes. Les résultats sur la tâche TREC Microblog 2012 n'ont cependant pas montré l'intérêt de ce facteur fraîcheur des tweets (Damak, 2014).

Ces résultats vont à l'encontre de la définition de la tâche de recherche TREC Microblog, ainsi qu'avec les résultats de l'état de l'art. Les mesures d'évaluation choisies pour la tâche et la façon dont les jugements de pertinence ont été effectués (sans tenir compte aucunement de la fraîcheur et de la nouveauté des informations) sont une explication plausible.

La tâche de filtrage de microblogs définie dans TREC 2015 met quant à elle en avant la nouveauté des tweets, mais cette fois en considérant des besoins utilisateurs récurrents (appelés profils). Nous avons proposé un modèle permettant de traiter le flux Twitter en temps-réel. Notre approche repose sur un processus de filtrage incrémental, dont le but est de séparer des tweets non pertinents le plus rapidement possible, sans dépasser quelques secondes entre l'arrivée d'un tweet et la décision finale pour un tweet passant toutes les étapes du filtrage (Chellal et al., 2015; Moulahi et al., 2016; Palmer et al., 2017). Une des

problématiques principales dans ce cadre, outre l'aspect temps-réel, est de ne pas surcharger l'utilisateur avec de l'information déjà connue ou inutile.

Nous avons défini une fonction de décision basée sur des seuils associés au contenu et au contexte du tweet. Outre la concordance du tweet à la requête, nous considérons les facteurs de contexte suivants :

— facteurs relatifs au contenu (nombre de termes réels, qualité du langage, qualité des hashtags) (Cheng et al., 2012; Duan et al., 2010) ;
— facteurs relatifs aux spécificités des tweets : nombre de hahstags, de mentions, présence d'URLs et d'images) (Nagmoti et al., 2010; Zhao et al., 2011) ;
— facteurs relatifs à l'auteur du tweet (nombre de followers, nombre d'amis, nombre de status, nombre de favoris...).

La plupart de ces facteurs ont été utilisés dans nos travaux précédents sur la recherche adhoc, à l'exception notable de certains facteurs relatifs aux auteurs des tweets, pour lesquels il n'était pas possible d'avoir l'information sur les collections 2011-2013. Pour chacun de ces facteurs, un seuil est défini, basé soit sur l'état de l'art, soit sur une analyse préalable de la collection (Palmer et al., 2017), permettant ainsi de passer outre le problème du démarrage à froid.

Nos premiers résultats, bien qu'encourageants en terme de vitesse (pas plus d'une seconde entre l'arrivée du tweet et la décision, même en période d'affluence), montrent que notre approche est pour le moment trop restrictive pour les requêtes/profils ayant un ensemble de résultats importants. Notre approche est en revanche performante pour les profils avec peu de tweets pertinents.

Parmi les pistes à envisager à la suite de ces travaux, nous pouvons citer l'ajustement dynamique des différents seuils utilisés, au fur et à mesure du processus (Zhao and Tajima, 2014). Jusqu'à présent, ils sont fixés en amont (issus d'expérimentations préalables) pour toute la durée de l'évaluation. Ils pourraient être recalculés et mis à jour après un certain laps de temps (une journée par exemple) à partir de l'analyse des données des précédentes exécutions. L'intégration d'un fenêtrage temporel (Gaglio et al., 2015; Zhao and Tajima, 2014) est également projetée. Il s'agit de découper la période de sélection de tweets en fenêtres de durées régulières pour récupérer les *tops* tweets par fenêtre et ensuite sélectionner parmi eux ceux qui seront transmis aux utilisateurs. Ceci permettrait d'éviter la transmission trop rapide d'un tweet alors que d'autres plus pertinents suivent juste derrière.

Un autre axe de recherche futur concerne le protocole d'évaluation utilisé dans les deux dernières campagnes d'évaluation TREC *Real-Time Summarization* (2016 et 2017). En analysant nos résultats sur la campagne 2016, nous avons trouvé un biais important dans le protocole d'évaluation, biais dont nous nous sommes servis pour notre participation en 2017 (Hubert et al., 2017a). Cela nous a permis d'obtenir respectivement les 2^{e} (évaluation temps-réel sur téléphone mobile) et 4^{e} place (évaluation *batch* plus traditionnelle) sur 40 participants. Je reviens plus en détail sur ces derniers travaux dans le Chapitre 5.

2.3 Synthèse des travaux présentés

Contributions principales. Nos contributions au domaine de la recherche dans les microblogs sont principalement expérimentales. Nos expérimentations ont été effectuées sur les collections présentées en Annexe A, Tableau A.3, ainsi que lors de participations officielles aux campagnes d'évaluation du domaine : TREC 2011 - *Microblog* (Damak et al., 2011), 2012 - *Microblog* (Ben Jabeur et al., 2012a), 2013 - *Microblog* (Ben Jabeur et al., 2013), 2015 - *Microblog* (Chellal et al., 2015), 2016 - *Real-time Summarization* (Moulahi et al., 2016) et 2017 - *Real-time Summarization* (Hubert et al., 2017b).

Elles ont mené à la mise en évidence de facteurs influençant la pertinence des tweets :

— Dans le cadre d'une tâche de recherche basée sur une requête et demandant un ensemble de tweets résultats, nous avons montré que les facteurs liés aux URLs et au contenu

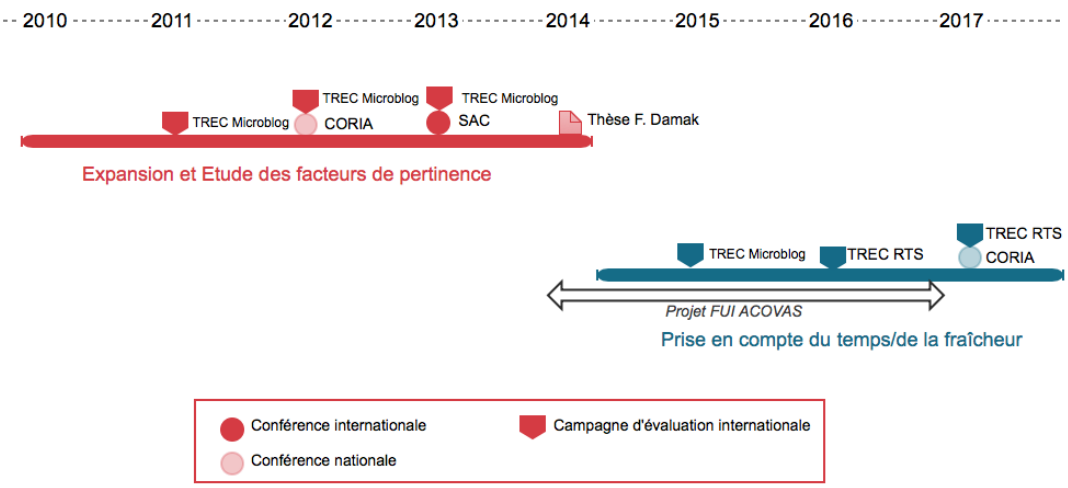


FIGURE 2.4 – Graphique synthétique de la structuration et de la valorisation des travaux sur le thème Recherche de microblogs

sont de grande importance pour la recherche. Parallèlement, certains facteurs de l'état de l'art tels que ceux liés aux hashtags, aux auteurs et à la notion de temps-réel n'ont pas d'intérêt particulier (Damak et al., 2012).

— Dans le cadre d'une tâche de filtrage temps réel des tweets pour répondre à un besoin utilisateur, tous les facteurs (relatifs au contenu, aux spécificités des tweets -hashtags, mentions, URLs- et aux auteurs) semblent par contre utiles (Palmer et al., 2017).

Ces dernières conclusions nous permettent de mettre en avant certaines limites des collections de test actuelles pour la recherche de microblogs. Jusqu'en 2015, les collections « statiques » proposées ne prenaient pas en compte la fraîcheur et la nouveauté des informations dans les jugements de pertinence, dénotant ainsi une inadéquation entre la tâche de recherche proposée et les besoins réels des utilisateurs. L'évolution de la tâche TREC *Microblog* vers la tâche *Real-time Summarization* a permis de tendre vers un cadre d'évaluation plus réaliste. Nous avons cependant découvert un biais important dans le nouveau protocole d'évaluation, biais dont l'analyse et la prise en compte nous a permis de nous classer 2^{e} sur les participations officielles en considérant l'évaluation temps-réel sur téléphone mobile.

Diffusion scientifique. La figure 2.4 synthétise de façon chronologique mes différentes publications citées dans le chapitre, organisées en fonction des différentes problématiques abordées.

Formation à la recherche. J'ai co-encadré avec Guillaume CABANAC et Mohand BOUGHANEM la thèse de Firas DAMAK, soutenue en 2014 (Damak, 2014). J'ai également co-encadré avec Gilles HUBERT la thèse de Thomas PALMER entre 2014 et 2017 [13].

Projets. Une partie de ces recherches a été menée dans le cadre du projet FUI Acovas (*Outils Agile pour la COnception et VAlidation Système*, 2014-2016), en partenariat (entre autres) avec les sociétés Airbus et Nexeya. Nous nous sommes intéressés dans ce projet à la détection d'anomalies dans des flux de données.

13. Thèse suspendue en Avril 2017.

CHAPITRE 3

Recherche de granules d'information autour des entités

Why is a raven like a writing desk?

— The Hatter.

Sommaire

3.1 État de l'art et problématiques ciblées

3.1.1 Introduction

Selon Pound et al. (2010), plus de la moitié des requêtes du Web ciblent une entité particulière ou des entités d'une classe. Une autre étude de la même année sur les *logs* des moteurs de recherche montre qu'entre 73% et 87% des requêtes contiennent des entités nommées et que jusqu'à 39% d'entre elles correspondent à exactement une entité nommée (Bautin and Skiena, 2009).

Pour répondre à ces requêtes orientées entités, des solutions particulières ont été proposées dans la littérature, comme alternative à la traditionnelle liste de documents. Avant de les présenter et de dégager les problématiques que nous avons abordées dans ce cadre, quelques définitions retenues dans nos travaux sont nécessaires.

Une **entité** est une « chose » qui peut être clairement identifiée (Chen, 1976). `Lewis Carroll`, `Wonderland` ou encore `the Hatter` sont des entités. Une **classe d'entités** est un ensemble d'entités du même type : `écrivains anglais`, `pays` ou encore `personnages de fiction` sont des classes d'entités. **Instance** ici est donc un synonyme d'entité nommée. Les deux possèdent une relation sémantique avec les **classes**. Un **attribut** d'une classe est une caractéristique de la classe. Par exemple, `date de naissance` est un attribut de la classe `écrivain`. Un attribut d'entité est un couple nom-valeur(s), la ou les valeurs pouvant être à leur tour des entités. Par exemple, l'entité `Lewis Carroll` peut avoir pour attributs `Date de naissance: 27/01/1832` ; `Lieu de naissance: Daresbury, Royaume-Uni` ; `Livres écrits: Alice in Wonderland, Through the looking glass` ; `Biographies : Lewis Carroll: A Biography, In the Shadow of the Dreamchild: A New Understanding of Lewis Carroll, The Mystery of Lewis Carroll`. Comme nous le voyons dans ces exemples, la valeur d'un attribut n'est pas forcément atomique et fixée. Dans les deuxième, troisième et quatrième cas, l'attribut a pour valeur des entités, dans les troisième et quatrième cas, l'attribut est multivalué et enfin dans le quatrième cas, les valeurs de l'attribut peuvent évoluer dans le temps.

En nous inspirant de (Alfonseca et al., 2010), nous distinguons trois types de besoin en information orientés entités :

— les requêtes **de type attribut** : `Date de publication d'« Alice in Wonderland »`, `Age d'Alice` ;

— les requêtes **de type entité** : `the Queen of Hearts`, `Through the looking glass` ;

— les requêtes **de type classe d'entités** : `Livres du 19e siècle`, `Habitants de Wonderland`.

En fonction du type de besoin, la réponse utilisateur (c'est-à-dire l'agrégat) peut différer :

— pour une requête par attribut, la seule valeur de l'attribut suffit (`1865` pour `Date de publication d'« Alice in Wonderland »` par exemple),

— pour une requête entité, on peut par exemple proposer un résumé de ses attributs les plus importants sous forme de tableau, ou encore proposer un résumé textuel des informations la concernant,

— pour une requête de type classe, le résultat peut être une liste des entités la composant, ou alors de façon plus complexe un tableau comparatif de ses différentes entités avec leurs valeurs respectives d'attributs.

Les travaux décrits dans ce chapitre se concentrent sur l'extraction/la recherche de granules pertinents autour des entités (deuxième partie du processus de recherche d'information agrégée, (re)voir Figure 2). L'agrégation de ces granules pour répondre finalement à la requête utilisateur est présentée dans le chapitre suivant. La frontière entre recherche de granules et agrégation étant parfois très mince dans ce cas, de nombreux liens existent entre nos deux chapitres (chapitres 3 et 4) .

Les problématiques liées à la recherche de granules documentaires autour des entités sont nombreuses, en témoigne la littérature abondante sur le sujet (Bautin and Skiena, 2009; Campos et al., 2015; Dietz et al., 2016; Ling et al., 2015). De notre côté, nous avons travaillé autour des problématiques suivantes :

— dans un premier temps, nous avons considéré l'entité et l'information liée comme fixes. Nous nous sommes intéressés à découvrir les relations liées aux entités, en considérant une relation comme un triplet (X, R, Y) (X et Y sont les entités, R la relation) ;

— dans un second temps, nous avons travaillé autour des entités pour lesquelles l'information bouge dans le temps, et avons cherché à extraire des documents vitaux les concernant.

3.1.2 Recherche de relations

Alors que la recherche d'information traditionnelle ne prend pas du tout en compte les relations, ces dernières sont étroitement liées aux requêtes orientées entités (requêtes de type attribut, entité ou classe). Tous les travaux de l'état de l'art sur l'annotation sémantique (*entity linking*), visant à lier des entités dans un texte à des ontologies ou des bases de connaissances, sont d'ailleurs basés sur ce lien étroit entre entités et relations (Shen et al., 2015). Comme nous l'avons vu plus haut, l'agrégat proposé à l'utilisateur pour ces requêtes orientées entités peut traduire les relations autour des entités :

— relation entité-classe, entre l'entité et sa classe (`The Dodo - is a - Alice in Wonderland fictional character`)

— relation entité-entité, entre deux entités. Il peut s'agir de relations spécifiques (`The King of Hearts- is married to- The Queen of Hearts`), ou de relations plus génériques (« est dans la même classe que », « est relatif à »).

— relation entité-attribut, entre une entité et ses attributs (`Lewis Carroll- is characterized by - Birth date`)

L'idée principale est que plus les informations sont mises en relation, plus le besoin en information sera satisfait.

La phase de recherche de granules documentaires peut donc être étendue dans ce cadre avec la recherche de relations. Les sources principales de contenu relationnel sont (i) le Web, (ii) les bases/graphes de connaissances et ontologies, et (iii) les bases de données relationnelles. On trouvera un état de l'art détaillé des approches proposées dans la littérature en fonction du type de relations visé dans (Kopliku et al., 2014). Nous synthétisons ici les approches de façon plus générique.

L'extraction de relations sur le Web est principalement réalisée à partir de méthodes d'extraction d'information, appliquées à des sites, pages ou parties de documents, afin d'extraire des classes, entités, attributs et leurs relations (Suchanek et al., 2009). La plupart des règles d'extraction sont de la forme $LxMyR$, où x et y sont des suites de termes, et L, M et R sont des *patterns* pouvant être trouvés respectivement avant, entre et après les deux groupes de termes. Par exemple, la règle « the x of y » peut être utilisée pour identifier les attributs des entités (`The birth date of Lewis Carroll is January 27, 1832`). Les règles d'extraction peuvent être apprises ou basées sur des heuristiques, elles utilisent des indicateurs variés comme des statistiques sur les termes, les balises, ou encore le *part-of-speech*. Les évidences sont combinées pour établir des règles correspondant aux classes, entités et attributs, ainsi qu'à leurs relations. Parmi les évidences les plus couramment utilisées on peut citer :

— les statistiques sur les termes (Agichtein and Gravano, 2000; Crescenzi et al., 2001; Etzioni et al., 2005) ;

— le *part-of-speech* (Bellare et al., 2007) ;

— les balises XML ou HTML (Cafarella et al., 2009; Crescenzi et al., 2001; Kopliku et al., 2011a) ;

— l'apparence visuelle (Aumann et al., 2006; Meng et al., 2003).

La plupart des approches d'extraction d'information, y compris celles citées au-dessus, sont spécifiques à certains types d'entités ou de classes. On notera cependant une exception pour les approches OIE (*Open Information Extraction*) initialement introduites par Banko et al. (2007). On remarque des efforts soutenus de la communauté autour de ces approches, qui permettent le remplissage automatique de bases de connaissances (Angeli et al., 2015; Etzioni et al., 2011; Fader et al., 2011; Lin and Etzioni, 2010; Zhu et al., 2009).

Les bases de connaissances telles que DBpedia, Freebase ou encore Yago, contiennent également une grande quantité de relations sémantiques (Limaye et al., 2010; Suchanek et al., 2008; Wu et al., 2008), et peuvent être utilisées pour apprendre les *patterns* d'extraction d'information ou encore pour renforcer des informations trouvées sur le Web.

Les techniques d'extraction d'information de l'état de l'art ne considèrent pas le problème de la recherche de relations comme un problème de recherche d'information, c'est-à-dire en partant d'une requête et en renvoyant des résultats en fonction de leur pertinence. Nous nous sommes, dans nos travaux, attachés à garder cet angle « Recherche d'information », et nous nous sommes focalisés sur deux types de relations particuliers :

— la relation entité-attribut : étant donnée une entité ou une classe d'entités, quels sont les attributs représentatifs ? en d'autres termes, quels sont les attributs les plus pertinents pour décrire ces entités ?

— la relation entité-entité : étant donnée une entité ou une classe d'entités, quelles sont les entités représentatives de la même classe ?

3.1.3 Filtrage temps réel de documents centrés sur une entité

Les bases de connaissances telles que Wikipedia, DBPedia ou FreeBase sont aujourd'hui les sources principales d'accès aux connaissances sur les entités (Li et al., 2012). 5,5 millions d'articles relatifs à des entités sont par exemple disponibles sur Wikipedia en 2018, selon les statistiques actuelles du site [1]. Toutefois, en raison du grand nombre d'articles Wikipedia, de nombreux contenus ne sont pas examinés par des experts, de sorte que le nombre de textes de mauvaise qualité a également augmenté considérablement (Suzuki and Yoshikawa, 2013). De plus, à cause du nombre limité de contributeurs (15 000) par rapport au nombre d'entités, signaler une nouvelle information en rapport avec une entité donnée sur sa page Wikipedia se fait généralement avec un temps de latence médian de 356 jours (Frank et al., 2012). Des problématiques similaires sont rencontrées sur les autres bases de connaissances. Pour obtenir des informations récentes sur les entités, ces sources ne peuvent donc être considérées comme des sources fiables, et des documents récents du Web doivent être trouvés. Les systèmes de filtrage du flux du Web peuvent répondre à cette problématique. On retrouve ces systèmes dans une variété d'applications (Abbes, 2015) : le suivi de célébrités (Zhou and Chang, 2013), le suivi d'événements en temps réel (Aslam et al., 2013a, 2015), la détection de nouvelles valeurs d'attributs des entités (Frank et al., 2014; Surdeanu and Ji, 2014), la mise à jour de bases de connaissances (Frank et al., 2014).

Lorsque l'on filtre des documents relatifs à une entité, deux scénarios possibles peuvent être identifiés :

— **Scénario 1 :** L'utilisateur ne connaît pas du tout l'entité. Dans ce cas, tous les documents qui parlent de cette entité pourraient l'intéresser. Nous nommons ces documents documents **centrés** sur l'entité.

— **Scénario 2 :** L'utilisateur connaît l'entité et veut mettre à jour ses connaissances. Dans ce cas, seuls les documents ajoutant une information nouvelle sur cette entité pourraient lui servir. Ces documents sont appelés documents **vitaux** (documents **frais**) dans la littérature. Les autres documents centrés sur l'entité mais répétant des informations déjà données dans le profil de l'entité sont **redondants**.

1. Source : http://en.wikipedia.org/wiki/Wikipedia:Size_of_Wikipedia , accédé le 23/3/2018.

Abbes (2015) a classé les différents critères utilisés dans les travaux de l'état de l'art pour filtrer les documents pertinents en quatre catégories :

— les critères a priori. Certains sont indépendants de l'entité, comme la longueur du document, sa source ou sa quantité d'information (Bouvier and Bellot, 2013; Wang et al., 2013). D'autres sont dépendants de l'entité mais indépendants des documents du flux. On peut par exemple citer le nombre d'entités reliées ou encore les types et les catégories des entités (Liu et al., 2013b; Wang et al., 2015a) ;

— les critères entité-document, qui peuvent être vus comme des fonctions à deux paramètres : le document et l'entité. En plus de la présence de l'entité et/ou de ses variantes dans le document, plusieurs critères basés sur le calcul de la similarité ou de la divergence entre le document et la page descriptive de l'entité (c'est-à-dire sa page Wikipedia) peuvent être exploités. On peut citer le cosinus, la similarité de Jaccard ou encore l'inverse de la KL-divergence (Wang et al., 2013, 2015a) ;

— les critères de pertinence temporels, qui visent à capturer s'il y a une augmentation brusque de l'intérêt autour de l'entité (on parle de **rafale**). Les rafales sont détectées en considérant un changement dans le volume de documents mentionnant l'entité (Balog and Ramampiaro, 2013) ou encore en utilisant des ressources externes telles que Google Trends [2] (Wang et al., 2015a). He et al. (2007) modélisent le flux par un automate à un nombre infini d'états dans lequel les rafales apparaissent comme des transitions d'états.

— les critères de pertinence basés sur des modèles de patrons (*patterns*). Jiang et al. (2014) ont par exemple remarqué que les documents vitaux contiennent souvent les entités en train d'effectuer certaines actions précises.

Ces critères sont ensuite utilisés par les différentes approches pour effectuer de la classification (Abbes et al., 2013b; Balog et al., 2013; Bonnefoy et al., 2013; Jiang et al., 2014) ou du tri de documents (Balog and Ramampiaro, 2013; Dietz and Dalton, 2013).

Si de nombreuses approches se sont penchées sur la détection de documents centrés sur les entités, la question de la vitalité des documents reste encore ouverte. Les approches de la littérature ne permettent pas de bien distinguer les documents vitaux des documents redondants. C'est sur cette dernière problématique que se sont penchés nos travaux.

3.2 Contribution au domaine de recherche

Nos contributions concernent à la fois la recherche de relations (section 3.2.1) et la recherche de documents vitaux (section 3.2.2).

3.2.1 Recherche de relations

3.2.1.1 Relations entité - attributs

Nous nous sommes dans un premier temps focalisés sur la recherche de relations entité - attributs, en cherchant les attributs représentatifs des entités. L'idée d'attribut représentatif est importante. Dans un contexte de recherche d'information, il ne s'agit pas de retrouver tous les attributs décrivant une entité ou une classe d'entités, au risque de se voir dépasser par le nombre d'attributs à considérer, mais plutôt les attributs les plus pertinents et intéressants pour l'utilisateur. Par exemple, pour la classe `écrivain anglais`, des attributs tels que `date de naissance` et `date de décès`, `lieu de naissance`, `bibliographie` auront probablement plus d'importance que `nom du père` et `nom de la mère`.

Afin d'extraire ces attributs représentatifs, nous avons considéré les tableaux HTML du Web (Kopliku et al., 2011a,d,e). Ces tableaux, très nombreux sur les pages Web, peuvent contenir des données relationnelles (Cafarella et al., 2008). Ils ont montré leur intérêt dans

2. http://www.google.com/trends, dernier accès en mars 2018.

	Year	Director
Alice in Wonderland	2010	Tim Burton
Alice Through the Looking Glass	2016	James Bobin

Tableau 1

Description	
Title	Alice in Wonderland
Director	Tim Burton
Year	2010
Release dates	
United Kingdom	February 25, 2010
United states	March 5, 2010

Tableau 2

1	Alice	Mia Wasikovsak
2	The Hatter	Johny Depp
3	White Queen	Anne Hathaway

Tableau 3

FIGURE 3.1 – Exemples de tableaux relationnels intéressants, exemple adapté de (Kopliku et al., 2011a)

la littérature pour l'extraction d'attributs (Chang et al., 2006; Chen et al., 2000). La tâche n'est cependant pas facile, puisque les tableaux HTML sont également utilisés à des fins de présentation ou de navigation. Cafarella et al. (2008) estiment que seuls 1% des tableaux HTML contiennent des données relationnelles. Un autre problème est que certains tableaux peuvent ne pas posséder d'entête, et donc de schéma (liste) d'attributs. Avant de sélectionner et d'attribuer un score aux attributs, notre approche commence donc par filtrer les tableaux pour ne conserver que ceux ayant un intérêt.

Filtrage des tableaux HTML. Notre approche consiste tout d'abord à interroger un moteur de recherche avec l'entité considérée. Les documents HTML renvoyés sont ensuite utilisés pour extraire des tableaux. Ces derniers sont très hétérogènes, et la plupart d'entre eux sont partiellement ou non pertinents. Afin de ne garder que les tableaux intéressants (voir Figure 3.1 pour des exemples de tableaux intéressants), nous appliquons les filtres suivants :

— un filtre relationnel et un filtre d'entête, qui visent à éliminer les tableaux non relationnels. Ces filtres sont basés sur des classifieurs orientés rappel et utilisant les caractéristiques définies dans (Cafarella et al., 2008), adaptées afin d'être également capables d'identifier des tableaux relationnels orientés horizontalement (par exemple les tableaux 1 et 3 de la figure 3.1 sont orientés verticalement et le tableau 2 est lui orienté horizontalement). Parmi les caractéristiques utilisées par le classifieur, on peut citer les dimensions du tableau, son remplissage, l'uniformité du texte pour le filtre relationnel, ou encore la conformité de l'entête et l'uniformité du texte par ligne pour le filtre d'entête ;

— un filtre de « ligne d'attributs », qui se place cette fois-ci au niveau des lignes/colonnes des tableaux et visant à identifier les lignes/colonnes contenant des noms d'attributs et des valeurs. Une ligne/colonne conforme doit contenir un nom d'attribut en première cellule et des valeurs d'attributs dans le reste des cellules. Ce filtre est également basé sur un classifieur et se base sur des caractéristiques telles que la présence de nombres, signes de ponctuation dans l'attribut potentiel, ou encore sa longueur en nombre de caractères (Kopliku et al., 2011e).

Score de pertinence. Une fois les attributs potentiels a identifiés pour une entité e, un score de pertinence $\phi(a, e)$ qui servira à les trier est ensuite calculé comme une combinaison de caractéristiques de pertinence :

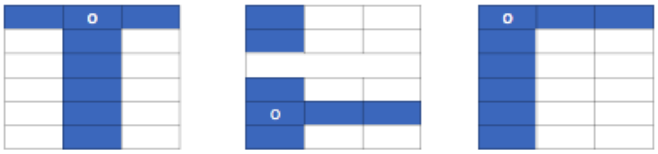


FIGURE 3.2 – Exemples de zones d'ombre pour une cellule de tableau O (Kopliku et al., 2011a)

$$\begin{aligned}\phi(a,e) = \quad & match(a,e,T) + drel(d,e) \\ & + log_{10}(search\_hits\_counts(a\ of\ e)) \\ & + DBPedia(a,e) + Wikipedia(a,e)\end{aligned} \tag{3.1}$$

Les opérandes de cette équation sont définies comme suit :

— $match(a,e,T)$ traduit le fait que les tableaux desquels les attributs sont extraits doivent être pertinents pour l'entité. Le fait d'être présent dans un document relatif à l'entité est nécessaire mais non suffisant. Soit a l'attribut extrait du tableau T pour l'entité e. L'appariement de e par rapport à une cellule du tableau $T_{x,y}$ est calculé comme le cosinus entre les termes de l'entité et les termes de la cellule. Le score du tableau est ensuite évalué comme suit :

$$match(e,T) = max_{T_{x,y} \in T}(cos(e, T_{x,y})) \tag{3.2}$$

Il est peu probable qu'un nom d'attribut apparaissent dans la même ligne/colonne que l'entité, contrairement à ses valeurs (on le voit clairement sur le tableau 1 de la figure 3.1). Pour tenir compte de ce fait, nous définissons une zone d'ombre pour chaque cellule O du tableau comme l'ensemble des cellules de la même ligne et de la même colonne, en enlevant éventuellement les cellules fusionnées (voir Figure 3.2). Nous avons au final :

$$match(a,e,T) = match(e,T) - match(e, shadow(a)) \tag{3.3}$$

avec $match(e, shadow(a)) = max_{T_{x,y} \in shadow(a)} cos(e, T_{x,y}))$.

— $drel(d,e)$ traduit la pertinence du document pour l'entité. L'idée est que plus un document est pertinent pour l'entité, plus les tableaux qu'il contient sont susceptibles de l'être.

$$drel(d,i) = \frac{\#results - rank}{\#results} \tag{3.4}$$

— $log_{10}(search\_hits\_counts(a\ of\ e))$ correspond au nombre de résultats renvoyés par le moteur de recherche à la requête « *attribute of entity* ». Cette caractéristique est présente dans la littérature pour l'extraction d'attributs (Popescu and Etzioni, 2005; Yoshinaga and Torisawa, 2007).

— $DBPedia(a,e)$ est égal à 1 si a est un attribut de e dans DBPedia (Auer et al., 2007), et $Wikipedia(a,e)$ est égal à 1 si a est présent dans l'*infobox* de la page Wikipedia de e. Ces deux facteurs sont évidemment à 0 si e n'est pas présent dans DBPedia ou Wikipedia.

Évaluation. Afin d'évaluer notre approche, nous avons construit un jeu de données composé de 200 entités réparties en 20 classes. On trouvera la liste de ces 20 classes sur la figure 3.3. Pour chacune des 200 entités, nous avons utilisé le top 50 des résultats fournis par l'API Yahoo Boss [3]. Trois problèmes de recherche différents ont été considérés :

3. https://developer.yahoo.com/boss/search/, dernier accès mars 2018. L'API n'est plus accessible depuis le 31/3/2016.

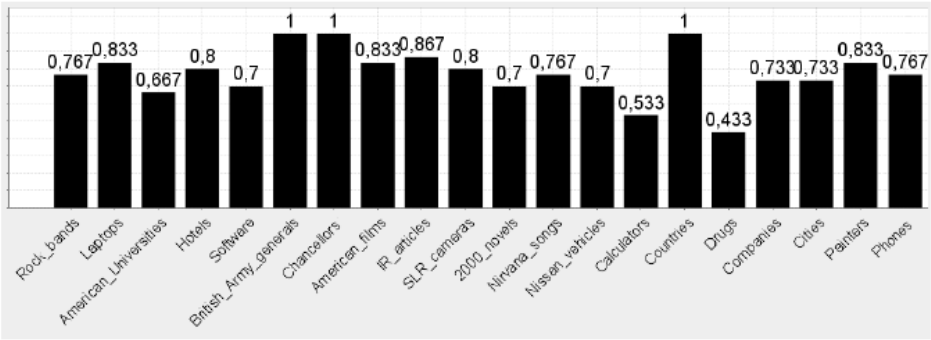


FIGURE 3.3 – Précision à 30 par classe d'entités (Kopliku et al., 2011a) (Q2)

— **Q1** : étant donnée une entité e, quels sont les attributs représentatifs ? Par exemple, quels sont les attributs représentatifs de l'entité `Lewis Carroll` ? Les attributs sont triés selon $\phi(a, e)$ (équation 3.2).

— **Q2** : étant donné un ensemble d'entités E, quels sont les attributs représentatifs de l'ensemble (c'est-à-dire de la classe) ? Par exemple, quels sont les attributs représentatifs des entités `Alice`, `King of Hearts`, `the Hatter`, `The White Rabbit` ? Les attributs sont alors triés selon l'équation suivante :

$$\phi(a, E) = \frac{\sum_{e \in E} \phi(a, e)}{|E|} \tag{3.5}$$

— **Q3** : étant donnés une entité et un ensemble $E+$ d'entités de la même classe, quels sont les attributs représentatifs de l'entité ? Par exemple, quels sont les attributs représentatifs de l'entité `Lewis Carroll`, sachant que `Charles Dickens`, `Oscar Wilde` ou encore `George Orwell` font partie de la même classe ? Les attributs sont alors triés selon l'équation suivante :

$$\phi'(a, e) = \begin{cases} 0 & \text{si } \phi(a, e) = 0 \\ \phi(a, e) + \phi(a, E+) & \text{sinon.} \end{cases} \tag{3.6}$$

Les résultats détaillés de ces expérimentations sont présentés dans (Kopliku et al., 2011a). Concernant **Q1** et **Q2**, nous obtenons un rappel élevé et une précision moyenne acceptable (voir Figure 3.3 pour les résultats sur **Q2**). Les résultats sont significativement améliorés sur **Q3**, ce qui montre que l'utilisation d'entités de la même classe est très utile pour retrouver les attributs représentatifs d'une entité. Notre approche a également été comparée aux approches utilisant des règles lexico-syntaxiques (Bellare et al., 2007; Paşca and Van Durme, 2008; Tokunaga et al., 2005; Yoshinaga and Torisawa, 2007), et obtient des résultats significativement meilleurs sur les précisions à 1, 10, 20 et 30 (61% contre 33% par exemple sur la précision à 30).

Intégration du Web de données. L'approche que nous venons de présenter possède les avantages d'être indépendante du domaine et de fonctionner pour toutes les entités, y compris celles qui ne sont pas présentes dans Wikipedia ou une base de connaissances. Le Web de Données est cependant une source importante de données structurées pouvant être utilisée pour notre problématique. Une autre limite de ces travaux est qu'ils se concentrent sur les noms d'attributs et laissent de côté la recherche des valeurs. Nous avons donc étendu nos travaux en évaluant la combinaison des deux sources d'évidence (Web et Web de Données) pour la recherche d'attributs et valeurs d'attributs représentatifs d'entités (Abbes et al., 2013a). Notre méthode de recherche dans les tableaux relationnels est simplement étendue pour récupérer toutes les valeurs des attributs présentes sur les lignes/colonnes des attributs représentatifs considérés. Pour ce faire, nous avons considéré 11 *datasets* du projet *Linking*

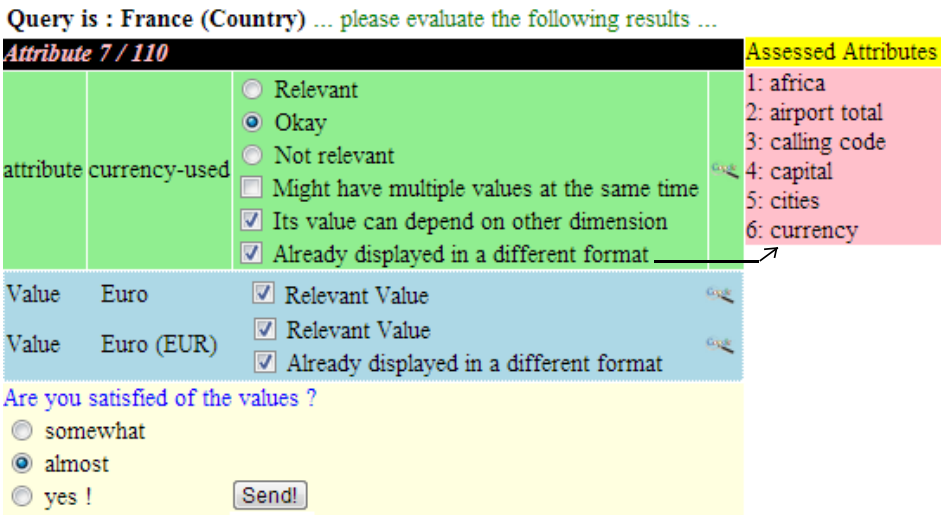


FIGURE 3.4 – Interface d'évaluation pour la recherche d'attributs représentatifs utilisant le Web et le Web de données (Abbes et al., 2013a)

Open Data : 2 datasets génériques (DBPedia et FreeBase) et 9 datasets spécifiques (aux domaines *Life Science*, *Government*, *Geographic*, *Media* et *Publications*). Ces datasets ont été interrogés avec des requêtes SPARQL, sur 57 requêtes entités réparties en 19 classes. Nous avons mis en place une interface permettant d'évaluer les résultats, dont on trouvera un aperçu sur la figure 3.4.

Les résultats, dont on trouvera le détail dans (Abbes et al., 2013a), montrent que les tableaux relationnels du Web sont très importants pour répondre à toutes les requêtes de type entité quelle que soit leur classe, mais leur précision reste faible par rapport au Web de Données. Des résultats similaires sont obtenus sur les valeurs des attributs, avec une précision à 73% pour les tableaux relationnels contre 90% pour le Web de données. La combinaison des deux sources semble donc une piste intéressante.

Un problème soulevé par ces expérimentations concerne les attributs dont les valeurs dépendent d'une autre dimension, comme le temps (la liste des livres écrits par Lewis Carroll a par exemple évolué au cours des années). Lorsque les informations relatives à une entité évoluent au cours du temps, la détection d'informations fraîches devient cruciale. Une partie de nos travaux s'est donc concentrée sur ce point, que nous détaillons en section 3.2.2.

3.2.1.2 Relations entités - entités

Connaître le type de relation entre entités est important pour la construction de l'agrégat final. Nous nous sommes pour notre part intéressés à un type particulier de relation, la relation « est de la même classe que » (Kopliku et al., 2010, 2011b). Pour extraire des entités de la même classe, il est commun dans la littérature de s'appuyer sur le nom de la classe (par exemple pour relier `Alice`, `the Hatter` and `the White Rabbit`, on commence d'abord par chercher la classe de ces entités, à savoir `personnages fictifs de Lewis Carroll`). Ceci peut être fait à partir de règles lexico-syntaxiques (Hearst, 1992) ou encore de bases de connaissances telles que DBPedia ou Wordnet (Auer et al., 2007; Hearst, 1998). Il n'est cependant pas toujours facile de trouver le nom de la classe, principalement parce qu'une entité peut appartenir à plusieurs classes : `Alice` appartient à la fois à la classe `personnages fictifs de Lewis Carroll` et `rôles dans des films de Walt Disney`), ou parce qu'il est difficile voire impossible de fixer une taxonomie des noms de classes (pour l'entité `Lewis Carroll` par exemple, la classe correspondante pourrait être `écrivains`, `écrivains anglais`, `écrivains du 19e siècle`). Notre approche se passe de cette étape d'acquisition de la classe, et cherche à extraire des groupes d'entités partageant la même classe en se basant sur les listes présentes dans les pages Web.

Nous avons mené une étude préalable sur un corpus composé de pages Web en français [4]. Sur les 2000 listes extraites et considérées, nous avons évalué que 8% d'entre elles étaient des listes d'entités partageant la même classe. Ce chiffre doit évidemment être confirmé sur un corpus plus étendu, mais montre que les listes du Web sont des sources intéressantes pour l'extraction d'entités de la même classe.

Afin de vérifier s'il était possible de les identifier sans trop de difficultés, nous avons construit plusieurs classifieurs SVM en nous basant sur des caractéristiques telles que la taille de la liste en nombre d'items et de caractères, la présence de liens ou encore la présence d'items redondants. Les résultats, bien que perfectibles, confirment l'intérêt des listes du Web pour l'extraction d'entités de la même classe.

3.2.2 Recherche de documents vitaux

Nous avons proposé deux approches principales pour le filtrage de documents vitaux :

— notre première approche est supervisée et se base sur les modèles de langue (Abbes et al., 2014c) ;

— notre seconde approche est non supervisée et combine deux facteurs : un facteur de pertinence thématique et un facteur de fraîcheur exploitant les dates reconnues dans le document (Abbes et al., 2015d,e).

Quelle que soit l'approche considérée, une phase préalable de filtrage est effectuée sur les documents. Cette phase détermine si les documents contiennent l'entité ou une variante de l'entité, et élimine les documents spams (Abbes et al., 2015e).

3.2.2.1 Modèles de langue pour l'identification de documents vitaux

Intuitivement, nous supposons qu'un document vital utilise un ensemble de termes qui peuvent refléter la vitalité. Par exemple, un document citant cette phrase `Johnny Depp vient d'accepter le rôle du Chapelier Fou dans le nouveau film de Tim Burton « Alice in Wonderland »`, a de fortes chances d'apporter une information nouvelle par rapport à un profil existant de l'entité `Johnny Depp`, puisqu'il utilise des termes pouvant refléter la vitalité comme `'vient'`, `'accepter'`, `'nouveau'`.

Dans notre approche, nous proposons de modéliser la notion de vitalité d'un document en nous basant sur les modèles de langue (Ponte and Croft, 1998). Nous proposons d'estimer un modèle de langue nommé modèle de vitalité, générant des documents vitaux par rapport à une entité donnée. Pour estimer le modèle vital d'une entité, nous supposons disposer d'un échantillon de documents vitaux pour cette entité. Nous notons cet ensemble vital $EV_e = \{dv_1, dv_2, \ldots, dv_m\}$.

dv_i représente tout le contenu ou une partie d'un document échantillon vital. Une seule partie de document (phrase, paragraphe, etc.) pourrait en effet être suffisante pour estimer sa vitalité. Dans la suite de l'article, par abus de langage, nous parlerons de document échantillon vital dv_i, dv_i représentant cependant tout ou une partie d'un document échantillon vital.

La vitalité peut être estimée de deux façons :

1. Elle peut être considérée comme unidimensionnelle. Nous pouvons alors estimer un seul modèle de vitalité θ_{Ve_u} à partir d'un seul document DV_e qui représente la concaténation de tous les documents échantillons vitaux de l'ensemble EV_e.
2. Elle peut être considérée comme multidimensionnelle. Nous pouvons dans ce cas estimer un sous-modèle de vitalité θ_{dv_i} pour chaque dv_i de l'ensemble EV_e.

4. Ce corpus a été constitué dans le cadre du projet européen Quaero.

Estimation d'un modèle de vitalité unidimensionnel. Étant donné un ensemble de documents échantillons vitaux EV_e pour une entité donnée, la probabilité de générer un terme t à partir d'un modèle de vitalité unidimensionnel θ_{Ve_u} est estimée de la façon suivante (lissage de Dirichlet) :

$$P(t|\theta_{Ve_u}) = \frac{tf(t, DV_e) + \mu\ P(t|C)}{|DV_e| + \mu} \tag{3.7}$$

où

— DV_e représente un document concaténant tous les documents échantillons vitaux appartenant à l'ensemble EV_e

— C est une collection de référence comportant tous les documents d'apprentissage et tous les documents de test.

— $tf(t, DV_e)$ représente la fréquence d'apparition du terme t dans le document DV_e

— μ représente une valeur de lissage réelle $\in [0, +\infty[$

— $P(t|C) = \frac{tf(t,C)}{\sum_{t' \in T} tf(t',C)}$, où T représente tous les termes du vocabulaire.

Estimation d'un modèle de vitalité multidimensionnel. Étant donné un ensemble de documents échantillons vitaux EV_e correspondant à une entité e composée de n termes, en faisant l'analogie par rapport au modèle de pertinence (Lavrenko and Croft, 2001), la probabilité de générer un terme t à partir d'un modèle de vitalité multidimensionnel θ_{Ve_m} est estimée comme suit :

$$\begin{aligned} P(t|\theta_{Ve_m}) &= \sum_{i=1}^{m} P(t|\theta_{dv_i}) P(\theta_{dv_i}|e) \\ &\propto \sum_{i=1}^{m} P(t|\theta_{dv_i}) P(e|\theta_{dv_i}) \end{aligned} \tag{3.8}$$

$$P(e|\theta_{dv_i}) = \prod_{t \in e} P(t|\theta_{dv_i}) \tag{3.9}$$

où

— $P(t|\theta_{dv_i})$ est l'estimation d'un modèle de vitalité unidimensionnel à partir d'un seul document vital dv_i. Cette probabilité est calculée selon l'équation 3.7.

— m représente le nombre de documents jugés vitaux pour l'entité e.

Mesure de la vitalité d'un document. Soit un nouveau document d composé de n termes t. Le score de vitalité du document d par rapport à un modèle de vitalité d'une entité θ_{Ve_x} (unidimensionnel θ_{Ve_u} ou multidimensionnel θ_{Ve_m}) est traduit par la vraisemblance des termes du modèle vital :

$$Score_{Vitalité}(d, e) = \prod_{t \in Ve_x} P(t|\theta_d)^{P(t|\theta_{Ve_x})} \tag{3.10}$$

où :

— $P(t|\theta_d)$ est calculé avec un lissage de Dirichlet,

— Nous considérons uniquement le *top* k des termes représentatifs dans le modèle θ_{Ve_x}.

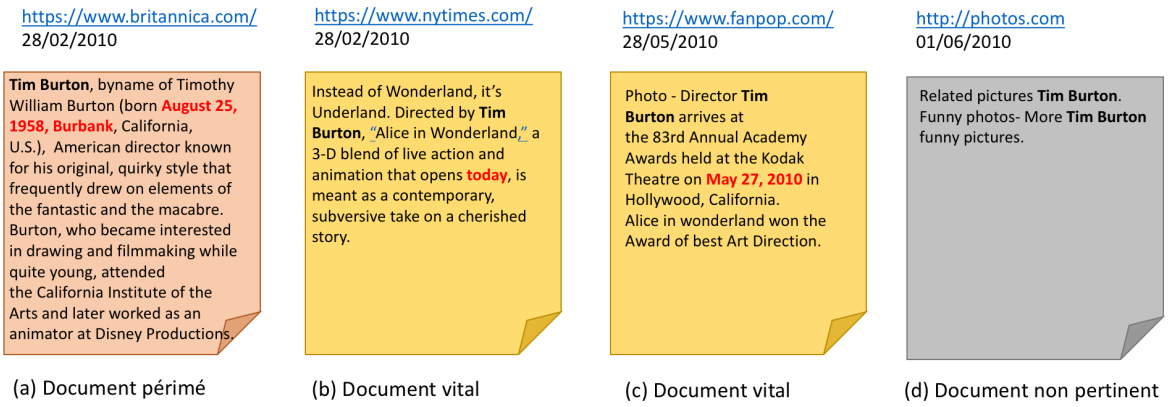


FIGURE 3.5 – Différence entre documents non pertinent, périmé et vital, en considérant la date de référence t_0 = Janvier 2010.

3.2.2.2 Exploitation des expressions temporelles pour la détection de la vitalité

Cette approche, non supervisée, repose sur l'idée qu'un document vital doit être récent. La fraîcheur d'un document peut être déterminée en regardant la date de publication du document et les expressions temporelles utilisées dans le texte. Nous prenons l'hypothèse qu'un document vital doit être récent (c'est-à-dire publié après une date de référence t_0) et qu'il doit contenir une date postérieure à t_0 et proche de sa date de publication. Par exemple, sur la figure 3.5, les documents (a) et (b) sont pertinents pour l'entité `Tim Burton` et ont été publiés tous les deux le `28 février 2010` . Le document (b) est plus frais que le (a) car il contient des expressions temporelles référant à une date (`today`, c'est-à-dire le `28 février 2010`) proche de la date de publication du document, alors que le document (a) contient une date plus ancienne (`25/08/1958`).

Soient une entité e et un document potentiellement vital d publié à la date $Date_p(d)$. Supposons que $Date_t^*$ est la date la plus proche de $Date_p(d)$ reconnue dans le document d. Nous supposons que plus le délai $\delta(d, E) = |Date_p(d) - Date_t^*|$ est court, plus grande est la probabilité de vitalité du document d. Formellement, nous évaluons le score de vitalité du document d par rapport à e de la façon suivante :

$$Score_{Vitalité}(d, e) = Pertinence(d, e) * (Fraîcheur(d, e) + \epsilon) \tag{3.11}$$

ϵ est destiné à éviter un score de vitalité nul lorsque le score de fraîcheur est égal à 0.

$$Fraîcheur(d, e) = e^{-\frac{\Delta(d,e)^2}{\sigma^2}} \tag{3.12}$$

$$\Delta(d, e) = \min_{x \in X(d,e)} (|Date_p(d) - Date_t(x, d)|^2) \tag{3.13}$$

— $Date_p(d)$ est la date de publication de d.

— $X(d, e)$ est l'ensemble des expressions temporelles détectées dans d (phrases, paragraphes, etc.) mentionnant E [5].

— $Date_t(x, d)$ est la date indiquée par l'expression x.

— $Date_t^*$ correspond à $Date_t(x, d)$ où $|Date_p(d) - Date_t(x, d)|$ est minimale. $Date_t^*$ doit être postérieure à la date de référence t_0, sinon le document est rejeté et est considéré comme périmé. La différence entre $Date_p(d)$ et $Date_t(x, d)$ est exprimée en jours.

5. Dans nos expérimentations, la détection d'expressions temporelles est effectuée grâce à la bibliothèque Sutime (Chang and Manning, 2012). Cette bibliothèque détecte aussi bien les dates complètes (comme dans le document (c) de l'exemple de la figure 3.5) que des expressions temporelles (comme dans le document (b)).

Le score *Pertinence* est utilisé pour prioriser les documents frais :

$$Pertinence(d,e) = \prod_{t \in top_k(P_e)} P(t|\theta_d)^{P(t|\theta_{P_e})} \tag{3.14}$$

— P_e est une page connue comme pertinente pour l'entité (par exemple sa page Wikipedia),

— $top_k(P_e)$ est l'ensemble des k termes les plus fréquents dans P_e,

— $P(t|\theta_d)$ et $P(t|\theta_{P_e})$ sont estimés en utilisant un lissage de Dirichlet comme décrit dans l'équation 3.15.

$$P(t|\theta_d) = \frac{tf(t,d) + \mu \frac{tf(t,C)}{\sum_{t' \in C} tf(t',C)}}{|d| + \mu} \tag{3.15}$$

— $tf(t,d)$ est la fréquence du terme t dans d

— $tf(t,C)$ est la fréquence du terme t dans la collection C

— C est la collection de référence composée des documents du flux antérieurs à t_0

— μ est un paramètre de lissage destiné à éviter les probabilités nulles.

3.2.2.3 Évaluation

Ces deux approches ont été évaluées sur les collections TREC KBA (*Knowledge Base Acceleration*) 2013 et 2014 (Frank et al., 2013, 2014) sur la tâche CCR (*Cumulative Citation Recommandation*). On trouvera le détail des expérimentations dans (Abbes et al., 2014b,c, 2015d,e). Nos conclusions sont les suivantes :

— Concernant l'approche par modèles de langue, chaque entité a son propre vocabulaire de vitalité, et la prise en compte de tout le contenu du document pour l'estimation du modèle permet d'obtenir de meilleurs résultats.

— Concernant notre deuxième approche, les dates à considérer doivent être proches de l'entité, et la considération des expressions temporelles dans des phrases permet une estimation plus précise de la vitalité.

Si l'on se compare à l'état de l'art, ces approches nous auraient permis de nous classer respectivement 1er et 2ème des tâches TREC 2013 et 2014. L'approche basée sur les expressions temporelles étant non supervisée, c'est celle que nous recommandons au final pour cette tâche.

3.3 Synthèse des travaux présentés

Les travaux présentés dans ce chapitre se sont intéressés à la détection de granules d'information liés aux entités. Plus précisément, nous nous sommes intéressés à extraire deux niveaux différents de granules d'informations :

— des attributs (et leurs valeurs) grâce à une méthode basée sur l'extraction de relations à partir des documents du Web. Cette méthode fonctionne pour n'importe quel type d'entité (Kopliku et al., 2011a,d,e), et il s'agit à notre connaissance d'une des premières approches de la littérature pour l'extraction d'attributs importants selon un angle Recherche d'Information.

— des documents vitaux, c'est-à-dire contenant de l'information nouvelle pour les entités variant dans le temps (Abbes et al., 2014c, 2015d,e). Les approches que nous proposons pour le filtrage de documents vitaux (approche par modèle de langue et approche basée sur les expressions temporelles) nous auraient permis de nous classer respectivement 1er et 2ème des tâches TREC *Knowledge Base Acceleration* 2013 et 2014.

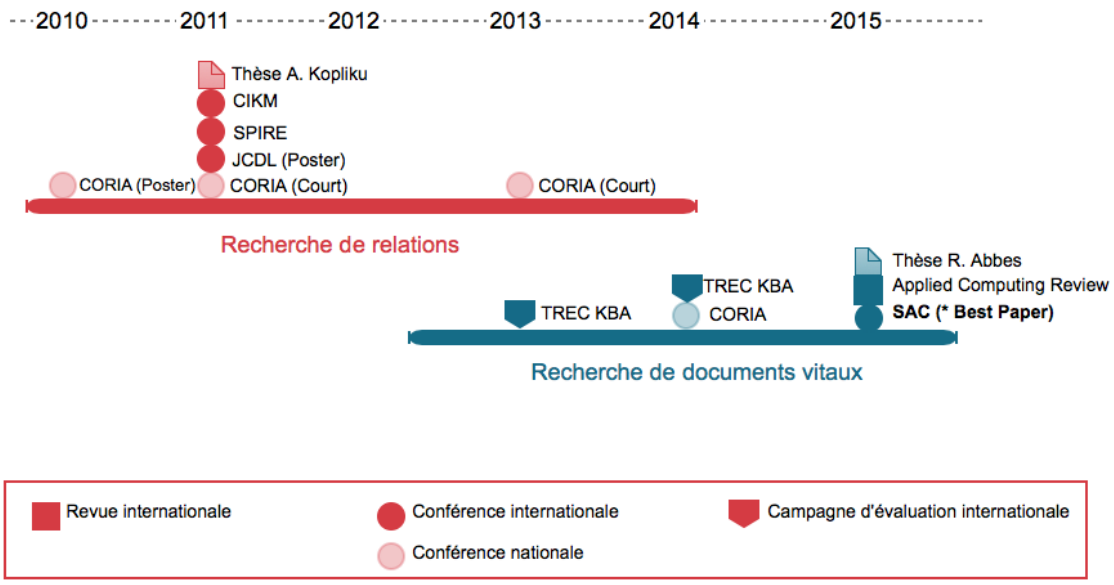


FIGURE 3.6 – Graphique synthétique de la structuration et de la valorisation des travaux sur le thème Recherche de granules autour des entités.

Nous avons effectué une évaluation de nos approches sur des collections construites par nos soins (recherche de relations) ou bien de l'état de l'art (TREC KBA 2013 et 2014, voir Annexe A, Tableau A.4). Notre approche pour la détection de documents vitaux nous a également permis de participer aux campagnes d'évaluation officielles TREC *Knowledge Base Acceleration* (KBA) 2013 (Abbes et al., 2013b) et 2014 (Abbes et al., 2014b).

Nos approches, nécessaires au processus d'agrégation, ne sont qu'une première étape de l'agrégation d'information autour des entités. Elles renvoient des granules pertinents, mais ne vérifient pas la redondance de l'information, pas plus qu'elles ne lient les agrégats. Cette étape finale de construction de l'agrégat est présenté dans le chapitre 4.

Diffusion scientifique. La figure 4.9 résume chronologiquement mes différentes publications relatives aux sujets abordés dans le chapitre.

Formation à la recherche. J'ai co-encadré un étudiant de Master sur ces thématiques (Abbes, 2012).

J'ai également participé à l'encadrement de deux thèses, dont une partie des contributions a été décrite dans le chapitre [6] :

— Arlind KOPLIKU (Kopliku, 2011), avec Mohand BOUGHANEM ;

— Rafik ABBES (Abbes, 2015), avec Mohand BOUGHANEM et Nathalie HERNANDEZ.

6. Pour ces deux thèses, la suite des contributions est présentée dans le chapitre 4.

CHAPITRE 4

Agrégation d'information autour des entités

> My dear, here we must run as fast as we can, just to stay in place. And if you wish to go anywhere you must run twice as fast as that.
>
> – The Queen of Hearts

Sommaire

4.1 État de l'art et problématiques ciblées

Ce chapitre se concentre sur la dernière étape du processus d'agrégation, à savoir la construction de l'agrégat (voir Figure 2 de l'introduction). Les travaux ayant abordé l'agrégation de résultats peuvent être classés selon deux catégories :

— L'**agrégation de sources** ramène le problème à un problème de fusion des résultats. Plusieurs sources d'information/de données sont interrogées, et les résultats sont ensuite fusionnés et renvoyés à l'utilisateur. Parmi les approches effectuant de l'agrégation de sources, on peut citer la meta-recherche (Selberg and Etzioni, 1995), la recherche fédérée (Callan, 2000), ou encore plus récemment la recherche agrégée inter-verticale (Arguello et al., 2009). Les moteurs de recherche actuels mettent tous en place la recherche d'information inter-verticale : des news, des images, des plans ou encode des vidéos sont présentés dans les résultats recherche.

— L'**agrégation de contenus** s'intéresse à construire un objet résultat à partir de granules d'information (textes, images, entités, attributs...) issus de différentes sources. Plusieurs sous-domaines de la recherche d'information se sont penchés sur le problème, parmi lesquels on peut citer la génération de langage naturel (NLG- *Natural Language Generation*) (Paris et al., 2010), les systèmes question-réponse (Moriceau and Tannier, 2010), ou encore le résumé de documents ou le résumé multi-document (Goldstein et al., 2000).

Cinq opérations basiques peuvent être considérées pour l'agrégation de contenus (Kopliku et al., 2014) :

— le **tri** : étant donné un ensemble de granules, l'opération de tri produit une liste triée des granules selon un critère déterminé (la pertinence, le temps,...) ;

— le **regroupement** : l'idée est de former des groupes de granules similaires ou complémentaires selon certains critères (contenu, même période temporelle, ou autre caractéristique commune). Des approches de *clustering* ou de classification peuvent être utilisées pour effectuer le regroupement ;

— la **fusion** (*merging*) : l'idée ici est de former un résultat cohérent à partir des granules. Contrairement au regroupement qui produit plusieurs groupes, un seul résultat sera produit : un nouveau document, un résumé, ou un objet (Nie et al., 2007).

— le **découpage** : c'est l'action opposée à celle de fusion. L'idée est de décomposer le granule en granules plus petits. Par exemple sur du texte, la décomposition peut fournir des passages, des phrases,...

— l'**extraction** : cette action, complémentaire à la précédente, a pour vocation d'extraire des granules qui peuvent être compris par eux-mêmes. De tels granules peuvent être des entités nommées, des images, des vidéos, etc.

Afin de former l'agrégat final, ces opérations peuvent être effectuées seules ou combinées (le tri peut par exemple être effectué sur des granules distincts ou sur des groupes de granules). L'idée derrière leur utilisation est de maximiser certaines propriétés des granules sur l'agrégat (non redondance, complémentarité, diversité, ...).

Nos travaux se sont plus particulièrement intéressés à l'agrégation de contenus [1]. Parmi les premières approches à proposer de l'agrégation de contenu, on peut citer (Basu Roy et al., 2010) dans laquelle les auteurs cherchent à relier des entités compatibles pour former un objet composite. Par exemple, un utilisateur souhaitant acheter un DVD du film `Alice au pays des merveilles` pour son enfant pourrait être intéressé par un agrégat contenant le DVD et une liste de produits dérivés proposés par `Disney`, le tout entrant dans son budget. La

1. L'agrégation de sources a été considérée sous un angle d'évaluation, nous revenons sur ce point dans le chapitre 5.

structure de l'objet ainsi que ses constituants sont connus à l'avance, la problématique ici consiste uniquement à trouver la meilleure combinaison qui satisfait les contraintes. Il n'y a cependant pas de problématique de recherche d'information. Ces travaux ont été étendus dans (Amer-Yahia et al., 2013), où l'objet composite est défini comme devant répondre à un certain nombre de contraintes (budget, complémentarité, cohésion et diversité). Dans les travaux cités précédemment, l'agrégation est assurée par le schéma qui guide la construction du résultat final. Afin de permettre une construction des objets composites à la volée, Bota et al. (2014) ont de nouveaux étendu les travaux de Basu Roy et al. (2010) en ajoutant la notion de pertinence à la requête. L'approche se concentre sur les requêtes informationnelles et compose l'agrégat à partir de verticaux du Web, en se limitant à deux types de critères de composition.

On retrouve également le concept d'agrégat dans des domaines spécifiques de la recherche d'information. En voici quelques-uns :

— les données politiques. Kaptein and Marx (2010) s'intéressent par exemple à des documents très longs contenant une transcription des débats au parlement hollandais. Au lieu de renvoyer les documents en réponse aux requêtes, les auteurs proposent un résumé des résultats avec un graphique à trois dimensions : l'année, le parti politique et le nombre de résultats. Les résultats peuvent ensuite être parcourus selon 3 facettes : par personne, par parti politique ou par année.

— la recherche académique, pour laquelle on peut citer des sites tels que Microsoft Academic Research [2] ou encore Google Scholar citations [3]. Cette fois, l'agrégation est principalement faite au niveau des chercheurs, en donnant leurs publications, leurs co-auteurs, ou encore des indicateurs globaux tels que le nombre de publications ou le *h-index*.

— les agrégateurs de news, dans lesquels les news avec des dates de publications et des sujets similaires sont regroupées dans des *news stories* (Rohr and Tjondronegoro, 2008; Sahoo et al., 2006), éventuellement présentées sous forme de *timeline* (Hennig and Wurst, 2006). Le site Google News [4] est un bon exemple d'agrégateur de news.

— les données géographiques. Dans ce cas, un plan est souvent juxtaposé avec une liste de résultats correspondants. Si la requête est suffisamment précise pour ne concerner qu'un seul lieu, d'autres contenus sont agrégés, comme des photos, des avis, des entités reliées, des articles de news (Kennedy and Naaman, 2008; Naaman et al., 2006; Vallet and Zaragoza, 2008). On trouvera sur la figure 4.1 un exemple de résultat sur le moteur de recherche Google Maps [5].

— les résumés d'opinions, dans lesquels on doit trouver les éléments clés pour les opinions positives ou négatives. Par exemple, à partir des critiques que l'on peut trouver sur le film `Les aventures d'Alice au pays des merveilles` de Tim Burton, l'idée est de présenter au spectateur potentiel un résumé des opinions positives et négatives, afin de l'aider à décider s'il regardera le film ou non (Kim et al., 2011).

Nous avons, pour notre part, abordé l'agrégation de contenus selon deux angles : la recherche agrégée relationnelle et le résumé temporel. La suite de cette section dresse un bref état de l'art pour chacun de ces axes de recherche et met en avant les problématiques que nous avons ciblées.

4.1.1 Recherche agrégée relationnelle

La recherche agrégée relationnelle peut être vue comme une généralisation de la recherche relationnelle (Cafarella et al., 2006) et de la recherche d'entités (Balog et al., 2009). L'agrégation ici est basée sur les relations entre les différents granules d'information (*nuggets*). En

2. http://academic.research.microsoft.com, dernier accès en mars 2018.
3. https://scholar.google.fr/citations, dernier accès en mars 2018.
4. http://www.news.google.com, dernier accès en mars 2018.
5. http://maps.google.org, dernier accès en mars 2018.

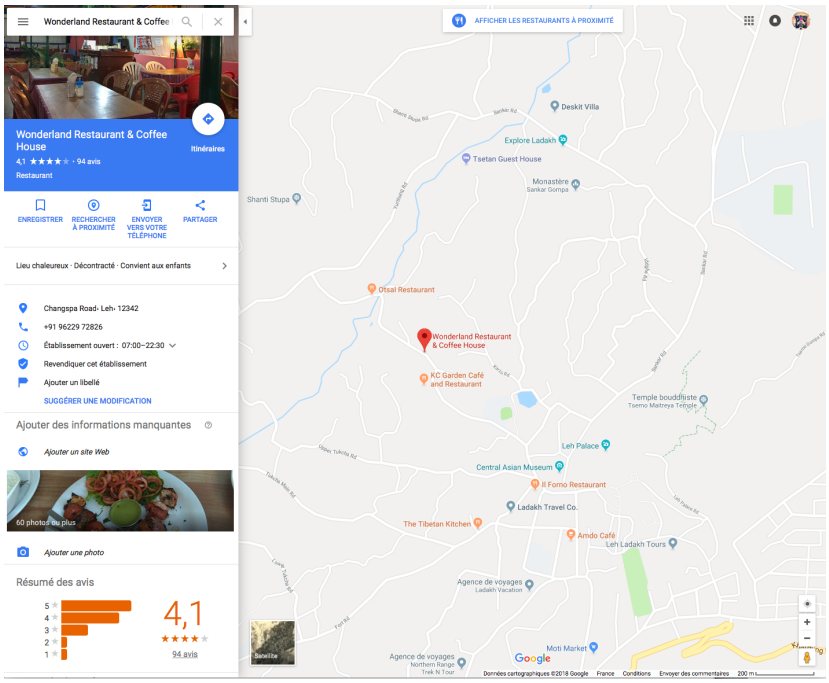


FIGURE 4.1 – Résultats de la requête `Wonderland restaurant` pour le moteur de recherche Google Maps, requête effectuée le 8/3/2018.

plus de la recherche de granules, la recherche agrégée relationnelle doit trouver des relations, nécessaires ensuite pour la phase d'agrégation des résultats. N'importe quel type de granules peut être utilisé : il est possible de mettre en relation des images, des vidéos, etc. Comme nous l'avons vu en détail au chapitre 3, il y a cependant 3 types de granules qui sont centraux au domaine : les classes, les entités et les attributs. De façon identique, toutes les relations *nugget-nugget* peuvent être intéressantes pour la recherche agrégée relationnelle. Parmi elles, la relation entité-nugget met en relation une entité avec n'importe quel type de contenu (par exemple `l'image X est une image du Chapelier Fou`). D'autre part, les granules classes, entités et attributs sont particulièrement intéressants car les relations entre eux sont communes et implicites. Nous avons identifié trois relations spécifiques : les relations entité-classe, entité-entité et entité-attribut.

Dans ce contexte et en fonction du besoin utilisateur, le type de l'agrégat final peut différer. Quand la requête est un attribut, le meilleur résultat est de retourner sa valeur. Quand la requête est une entité, le meilleur choix pourrait être un résumé des différents attributs. Quand la requête est une classe, le résultat pourrait être un tableau comparatif des entités de cette classe et de leurs attributs.

La séparation entre recherche de granules et de relations et l'agrégation de résultats est dans ce contexte parfois mince. Nous nous sommes focalisés au chapitre 3 sur la recherche de relations et plus particulièrement de relations entités-attributs et entités-entités, nous nous focalisons maintenant sur la construction de l'agrégat.

Parmi les approches effectuant de la recherche agrégée relationnelle, nous pouvons citer le travail de (Cafarella et al., 2006) qui a mis en place un moteur de recherche intitulé TextRunner. Ce moteur essaie de répondre à des requêtes complexes qui exigent comme résultat une liste d'entités ou bien un tableau en exploitant un graphe d'entités-relations. En premier lieu, l'approche proposée cherche à traiter la liste d'entités. Pour ce faire, un robot d'exploration standard est d'abord utilisé pour récupérer des pages web. Par la suite, les auteurs appliquent un mécanisme d'extraction d'information pour chaque phrase sur chaque page récupérée. La plupart des informations extraites sont soit des objets (entités) soit des prédicats (relations). Ces derniers vont être reliés pour y faire un graphe G. Afin de faciliter le traitement des requêtes, TextRunner calcule un index inversé sur le texte des triplets

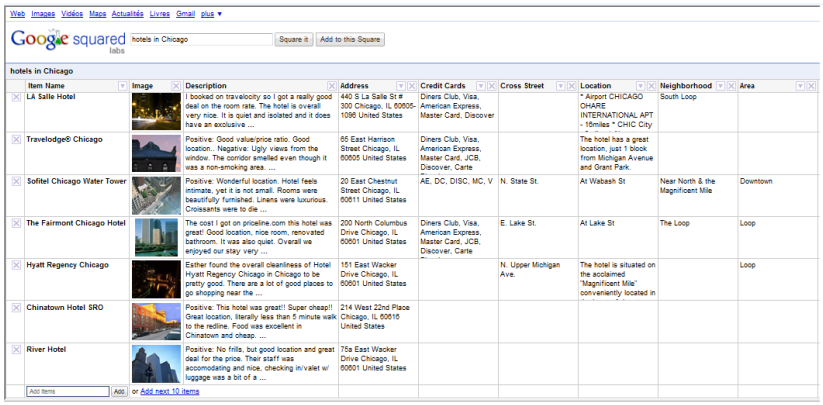


FIGURE 4.2 – Résultats de la requête `Hotels in Chicago` pour le moteur de recherche Google Squared, 2010.

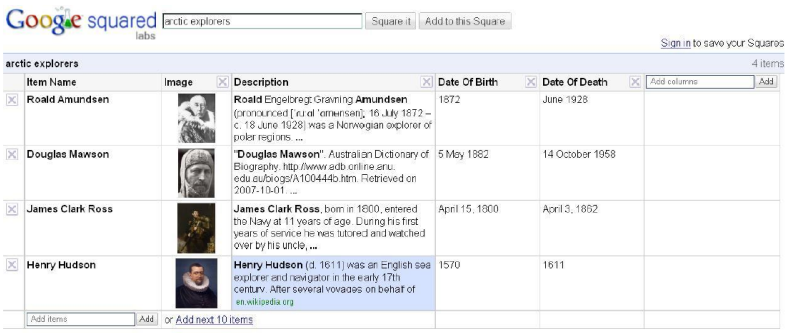


FIGURE 4.3 – Résultats de la requête `Artic explorer` pour le moteur de recherche Google Squared, 2010.

qui forment le graphe d'extraction G. L'évaluation a prouvé l'efficacité de cette méthode à retrouver les bonnes entités par rapport aux systèmes d'extraction d'information qui se basent sur des *patterns* classiques.

De son côté, GoogleLabs [6] a lancé en 2009 un outil expérimental, intitulé *Google Squared*, qui permettait de générer un tableau descriptif pour une requête donnée de type classe (voir deux exemples sur les figures 4.2 et 4.3). Google Squared offrait plusieurs fonctionnalités, comme la possibilité de modifier le tableau en retirant les lignes et les colonnes inutiles, ou d'ajouter de nouvelles lignes et colonnes en lui laissant rechercher les faits pertinents. Malgré son originalité, cette dernière approche retournait souvent des tableaux troués et renvoyait parfois des attributs avec des valeurs erronées.

Si l'information retrouvée par Google Squared était extraite du Web, Wolfram Alpha [7], introduit également en 2009, propose quant à lui une approche différente où les faits relationnels sont extraits d'une base de connaissance interne pré-construite (voir Figure 4.4).

Le *Google Knowledge Graph* introduit en 2012 par Google est un autre exemple d'utilisation de faits relationnels (Singhal, 2012). Le graphe est construit en utilisant à la fois des bases de connaissances telles que Freebase ou Wikipedia, mais aussi le Web dans son ensemble. En 2014, Google annonçait 1.6 milliard de faits indexés, dont 271 millions de « certains », c'est-à-dire sûrs à plus de 90% [8]. Le *Google Knowldege Graph* est utilisé par Google pour désambiguïser les résultats, présenter des résumés sur certaines entités nommées (voir par exemple la Figure 4.5), et pour aider l'utilisateur à aller plus loin dans sa recherche en

6. http://www.googlelabs.com : le site GoogleLabs de Google est actuellement fermé, et il n'est malheureusement plus possible d'accéder à GoogleSquared depuis Septembre 2011 (https://en.wikipedia.org/wiki/Google_Squared, accédé le 5/3/2018.).

7. http://www.wolframalpha.com/, dernier accès en mars 2018

8. Source : Wikipedia, https://en.wikipedia.org/wiki/Knowledge_Graph, accédé le 6/3/2018.

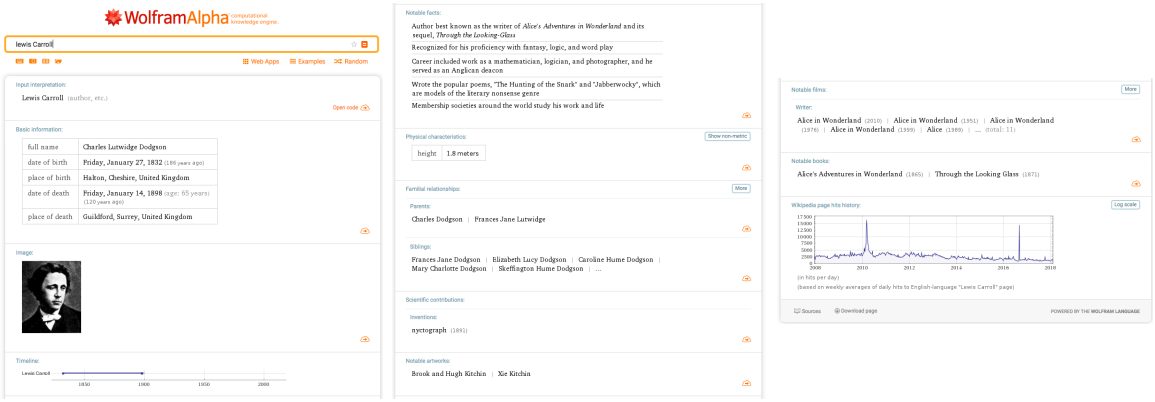


FIGURE 4.4 – Résultats de la requête `Lewis Carroll` pour le moteur de recherche Wolfram Alpha, requête effectuée le 4/3/2018.

FIGURE 4.5 – Résultats de la requête `Lewis Carroll` pour le Google Knowledge Graph, requête effectuée le 4/3/2018.

lui présentant de nouveaux faits/relations en rapport avec sa recherche.

Une autre représentation basée sur des tableaux a été proposée par Elmeleegy et al. (2009). Les tableaux sont extraits à partir de listes. Tout d'abord, de multiples sources d'informations sont utilisées pour segmenter les différentes lignes d'une liste en de multiples champs. Ensuite,

selon le nombre de segments obtenus dans la majorité des lignes, le système décide du nombre de colonnes à générer dans le tableau et réorganise les mauvais segments. Finalement, en exploitant les tableaux du web, le système affine le tableau obtenu en remplissant les cases vides par des valeurs correspondantes. Pour chaque table extraite, les auteurs calculent un score d'extraction qui reflète la qualité de la table produite.

Nous nous sommes pour notre part focalisés sur l'assemblage des entités et des attributs pour répondre à des requêtes classe. L'agrégation est en effet plus ardue dans ce cas. Pour les requêtes attribut, il « suffit » de renvoyer des valeurs candidates triées par pertinence. Pour les requêtes entité, il faut retourner des attributs pertinents, mais aussi sélectionner quels sont les attributs les plus représentatifs. Pour les requêtes classe le problème est encore plus complexe, puisqu'il faut **sélectionner et assembler les entités et attributs importants**. L'ordre utilisé pour présenter les entités et les attributs importants peut affecter la qualité du résultat agrégé. Ce dernier peut prendre différentes formes, la plus intuitive étant celle du tableau, que nous avons utilisée dans nos travaux.

4.1.2 Résumé temporel

Les approches que nous avons décrites dans le chapitre 3 filtrent les documents centrés sur une entité, en ne conservant que ceux qui contiennent de l'information vitale, c'est-à-dire de l'information nouvelle. Un utilisateur intéressé par ces informations nouvelles devra cependant parcourir tous les documents pour extraire les faits nouveaux, et éliminer les redondances potentielles. En effet, les documents renvoyés peuvent contenir des informations équivalentes : la redondance de l'information n'a pas du tout état considérée dans notre approche.

Agréger les informations nouvelles extraites des documents vitaux revient à construire un résumé temporel. Un résumé idéal est court, sans redondance, et le plus exhaustif possible sur l'entité d'intérêt.

Plusieurs applications sont possibles :

— la mise à jour de bases de connaissances (Frank et al., 2013; Surdeanu and Ji, 2014) ;

— le résumé temps réel de microblogs. Les granules (par exemple les tweets) sont extraits de plateformes de microblogging. Les microblogs étant caractérisés par une faible longueur, les granules extraits ici sont agrégés tels quels pour former le résumé. Nous avons mentionné cette application dans le chapitre 2 ;

— le résumé temporel autour d'entités. L'idée est de produire une synthèse autour d'une entité donnée. Cela peut être particulièrement utile par exemple lorsque les entités sont des événements (par exemple des catastrophes naturelles, accidents ou attentats) pour lesquels un résumé des évolutions est parfois crucial pour les utilisateurs, surtout s'ils sont directement concernés (Aslam et al., 2013a).

Le résumé temporel de documents se base sur des approches issues du résumé automatique de documents (Nenkova et al., 2011) ou encore s'intéressant à la détection de la nouveauté (Soboroff and Harman, 2005). Si l'on regarde l'état de l'art, de nombreuses approches ont été proposées dans le cadre de la tâche TREC *Temporal Summarization* (Aslam et al., 2013a, 2015), active de 2013 à 2015.

Les approches proposées suivent généralement un processus en deux phases :

— une 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 (Zhao et al., 2014), le TF.IDF (Xu et al., 2013), le modèle booléen appliqué sur les titres des documents d'actualité (Liu et al., 2013a) ou sur tout le contenu des documents (Xu et al., 2013).

— une 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 des mots importants relatifs à l'événement qui se trouvent dans la phrase (Chen et al., 2014; Liu et al., 2013a; Zhang et al., 2013)

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. (2013) 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. (2014) ont proposé un modèle de régression exploitant plus de 300 critères décrivant les aspects prévalence, nouveauté et qualité des phrases.

Nous nous sommes, pour notre part, focalisés sur les trois questions de recherche suivantes :

— À 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 ?

— À 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 ?

4.2 Contributions au domaine de recherche

Comme indiqué précédemment, nos recherches sur l'agrégation de contenus sont ici présentées selon deux axes : (i) la recherche agrégée relationnelle et la construction de résultats tabulaires et (ii) la construction de résumés temporels.

4.2.1 Recherche agrégée relationnelle

Les approches présentées dans le chapitre 3, section 3.2.1 nous permettent de retrouver des relations entités - attributs et entités - entités. Une fois les relations détectées, se pose évidemment le problème de la construction de l'agrégat. Comme nous l'avons vu dans la section 4.2.1, les tableaux sont une forme relativement commune de présentation des résultats. La suite logique des travaux présentés au chapitre 3 sur la recherche d'attributs importants était donc de présenter ces attributs importants et leurs valeurs dans un tableau. La figure 4.6 donne un exemple d'agrégat pour 3 entités, basé sur l'approche proposée pour répondre à **Q2** (voir section 3.2.1).

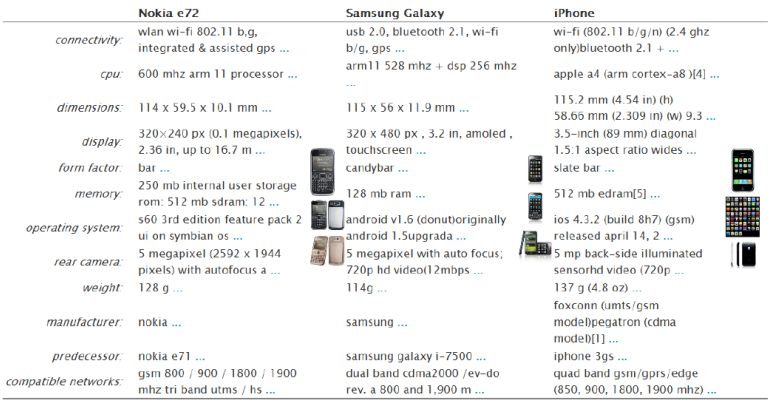
Les travaux que nous présentons maintenant sont très liés aux précédents, mais nous nous sommes focalisés ici sur les requêtes classe. Afin d'illustrer le problème de l'agrégation de résultats pour des requêtes classe, nous avons construit quelques exemples présentés sur la figure 4.7. Ils correspondent à des résultats possibles pour la requête `Ecrivains anglais du 19ème siècle`. Chaque ligne correspond à une entité et chaque colonne à un attribut.

Les 3 premiers tableaux contiennent des résultats non pertinents :

— le tableau A a des attributs non pertinents (par exemple : `Email`) et des attributs peu pertinents (`Collège`, `Nationalité`) ;

— le tableau B contient une entité non pertinente (`Victor Hugo`) et des écrivains peu représentatifs (`Hall Caine, Frederic Farrar`) ;

— les problèmes dans le tableau C viennent des valeurs des attributs. La plupart d'entre elles sont manquantes.



	Nokia e72	Samsung Galaxy	iPhone
connectivity:	wlan wi-fi 802.11 b,g, integrated & assisted gps ...	usb 2.0, bluetooth 2.1, wi-fi b/g, gps ...	wi-fi (802.11 b/g/n) (2.4 ghz only)bluetooth 2.1 + ...
cpu:	600 mhz arm 11 processor ...	arm11 528 mhz + dsp 256 mhz ...	apple a4 (arm cortex-a8)[4] ...
dimensions:	114 x 59.5 x 10.1 mm ...	115 x 56 x 11.9 mm ...	115.2 mm (4.54 in) (h) 58.66 mm (2.309 in) (w) 9.3 ...
display:	320×240 px (0.1 megapixels), 2.36 in, up to 16.7 m ...	320 x 480 px , 3.2 in, amoled , touchscreen ...	3.5-inch (89 mm) diagonal 1.5:1 aspect ratio wides ...
form factor:	bar ...	candybar ...	slate bar ...
memory:	250 mb internal user storage rom: 512 mb sdram: 12 ...	128 mb ram ...	512 mb edram[5] ...
operating system:	s60 3rd edition feature pack 2 ui on symbian os ...	android v1.6 (donut)originally android 1.5upgrada ...	ios 4.3.2 (build 8h7) (gsm) released april 14, 2 ...
rear camera:	5 megapixel (2592 x 1944 pixels) with autofocus a ...	5 megapixel with auto focus; 720p hd video(12mbps ...	5 mp back-side illuminated sensorhd video (720p ...
weight:	128 g ...	114g ...	137 g (4.8 oz) ...
manufacturer:	nokia ...	samsung ...	foxconn (umts/gsm model)pegatron (cdma model)[1] ...
predecessor:	nokia e71 ...	samsung galaxy i-7500 ...	iphone 3gs ...
compatible networks:	gsm 800 / 900 / 1800 / 1900 mhz tri band utms / hs ...	dual band cdma2000 /ev-do rev. a 800 and 1,900 m ...	quad band gsm/gprs/edge (850, 900, 1800, 1900 mhz) ...

FIGURE 4.6 – Agrégat formé avec les attributs importants de 3 entités téléphone (Kopliku et al., 2011a).

	Naissance	Décès	Email	Collège	Nationalité
Jane Austen	1775	1817		Oxford	Anglaise
Arthur Conan Doyle	1859	1930	Arthur.conan.doyle@gmail.com	Hodder Place	Anglaise
Lewis Carroll	1832	1889		Rugby school	Anglaise
Charles Dickens	1812	1870	Charles.dickens.33@yahoo.com	/	Anglaise

Tableau A

	Naissance	Décès	Œuvre la plus lue	Œuvre la plus adaptée au cinéma
Victor Hugo	1802	1885	Ruy Blas	Les misérables
Lewis Carroll	1832	1889	Alice au pays des merveilles	Alice au pays des merveilles
Hall Caine	1853	1931	The Deemster	The Bondman
Frederic Farrar	1831	1903	An Essay of the Origin of Language	/

Tableau B

	Œuvre la plus lue	Œuvre la plus adaptée au cinéma	Genre
Jane Austen	Orgueils et préjugés	Orgueils et préjugés	Nouvelles
Arthur Conan Doyle		Le chien des Baskerville	Policier
Lewis Carroll	Alice au pays des merveilles		
Charles Dickens			Drame

Tableau C

	Naissance	Décès	Œuvre la plus lue	Œuvre la plus adaptée au cinéma	Genre
Jane Austen	1775	1817	Orgueils et préjugés	Orgueils et préjugés	Nouvelles
Arthur Conan Doyle	1859	1930	Les aventures de Sherlock Holmes	Le chien des Baskerville	Policier
Lewis Carroll	1832	1889	Alice au pays des merveilles	Alice au pays des merveilles	Nouvelles
Rudyard Kipling	1865	1936	Tu seras un homme, mon fils	The jungle book	Nouvelles, Poésie
Charles Dickens	1812	1870	David Copperfield	Un chant de Noël	Drame

Tableau D

FIGURE 4.7 – Exemples de résultats tabulaires pour la requête classe `Ecrivains anglais du 19ème siècle`. Adapté de (Krichen et al., 2012).

Le tableau D semble être le plus pertinent : il contient des entités (représentatives) de la classe et des attributs importants et pertinents, avec leur valeur. La qualité des résultats de recherche agrégée dépend de la qualité des entités, des attributs et de leur valeur. Idéalement, les entités et les attributs ne doivent pas être seulement pertinents mais aussi représentatifs pour la requête (ici la classe), et les valeurs d'attributs doivent être présentes et correctes.

Afin d'agréger les résultats d'une requête classe, une première étape est de rechercher les entités et les attributs liés à la classe. Il est ensuite nécessaire de *(i)* sélectionner les entités représentatives de la classe, et *(ii)* trier ses attributs. Nous avons considéré la première étape comme acquise (sélection des nuggets), et nous sommes focalisés sur les deux derniers points [9].

9. Il est à noter que la sélection des entités/attributs représentatifs dans le cas des requêtes classe aurait également pu être abordé au chapitre précédent, comme nous l'avons fait pour la section des attributs importants pour les requêtes entités. Nous avons cependant fait le choix de le détailler ici, l'approche présentée se focalisant également sur la construction de l'agrégat, c'est-à-dire du tableau résultat. La frontière entre acquisition des granules/relations et construction de l'agrégat est parfois très mince.

4.2.1.1 Tri des entités

Afin de classer les attributs les plus représentatifs d'une classe, nous sommes partis de l'hypothèse suivante : les attributs représentatifs doivent être présents dans les entités « importantes ». Le tri des attributs est donc très lié à celui des entités : qu'est-ce qu'une entité « importante » ? Dans le but d'analyser l'impact des entités sur le classement des attributs, nous avons mis en place trois algorithmes de sélection d'entités :

— le premier, intitulé *Shorty*, sélectionne les n premières entités qui ont le moins d'attributs parmi celles qui ont au moins un attribut. L'idée est que le peu d'attributs présents sont forcément importants.

— le deuxième algorithme, intitulé *Maxy*, sélectionne les n premières entités qui ont le plus d'attributs. L'idée cette fois est que les attributs importants sont forcément présents dans le nombre.

— finalement le troisième algorithme, intitulé *Fantasy*, part de l'hypothèse que les entités bien référencées (donc populaires) sont importantes, et utilise certains résultats de l'état de l'art sur la recherche d'entités nommées (Balog et al., 2009; Vercoustre et al., 2007). Au niveau de chaque entité de la classe c, il existe des valeurs d'attributs, elles-mêmes de type entité, qui pointent vers d'autres entités et inversement (des entités qui pointent vers l'entité de la classe c à travers des valeurs d'attributs). Plus concrètement, nous sélectionnons les entités en fonction du nombre de liens entrants ($interLink$) vers la page associée de Wikipedia et du nombre de liens sortants ($extLink$) de cette page.

4.2.1.2 Tri des attributs

Nous devons, à ce niveau, trier les attributs en calculant un score $score\,(a)$ pour chaque attribut a de la classe c servant de requête afin d'identifier les attributs les plus pertinents pour l'agrégation. Cette étape permet à la fois de mettre en valeur les attributs importants et de minimiser le risque de valeurs non renseignées dans le tableau. Nous proposons de calculer le $score\,(a)$ de trois façons différentes.

— **Score basé sur la fréquence (1ère formule)**
En premier lieu, nous proposons d'utiliser la fréquence d'apparition des attributs dans la classe traitée ($tf\,(a, c)$) sur le nombre d'attributs dans la classe en question pour normaliser ($|c|$).

$$score\,(a) = \frac{tf\,(a, c)}{|c|} \tag{4.1}$$

Pour cette formule nous ne prenons pas en compte le classement des entités et considérons que toutes les entités ont le même impact.

— **Score basé sur la probabilité de pertinence (2e et 3e formules)**
Une autre façon de faire est de considérer le score $score\,(a)$ comme une probabilité d'appartenance $P\,(a|c)$ à la classe recherchée c pour chaque attribut traité a.
En traitant les attributs comme des termes, les entités comme des documents et la classe comme la collection traitée, nous pouvons faire une analogie avec le principe de classement probabiliste (*Probability Ranking Principle*), énoncé par Robertson (1997). Dans notre cas, le mieux est de retourner les attributs a et les entités i en ordre décroissant de leur probabilité sachant la classe c.

Dans la suite, nous allons utiliser la formule de probabilité totale de cette façon :

$$score\,(a) = P\,(a|c) \ = \sum_{e \in c} P\,(a|e) \ \cdot P\,(e|c) \tag{4.2}$$

avec :

— $P\,(a|e)$: probabilité de pertinence de l'attribut a sachant l'entité e.

— $P(e|c)$: probabilité de pertinence de l'entité e sachant la classe c.

La probabilité $P(a|i)$ peut être calculée selon les deux formules suivantes :

— **Deuxième formule**
$P(a|e)$ est égale à l'inverse du nombre d'attributs $|e|$ dans l'entité e.

$$\begin{aligned} score(a) = P(a|c) \quad &= \sum_{e \in c} (P(a|e) \cdot P(e|c)) \\ &= \sum_{e \in c} \left(\frac{1}{|e|} \cdot P(e|c) \right) \end{aligned} \tag{4.3}$$

Tous les attributs ont le même poids au sein de l'entité à laquelle ils appartiennent mais les attributs des entités ayant le moins d'attributs deviennent plus représentatifs que ceux des autres entités.

— **Troisième formule**
Pour cette troisième formule, nous utilisons le ratio de la fréquence $tf(a, c)$ de l'attribut a dans la classe c sur la somme des fréquences $tf(a_j, c)$ des attributs a_j de la classe c à laquelle appartient l'entité e en question. Ce qui donne comme formule finale :

$$\begin{aligned} score(a) = P(a|c) \quad &= \sum_{e \in c} (P(a|e, c) \cdot P(e|c)) \\ &= \sum_{e \in c} \left(\frac{tf(a,c)}{\sum_{a_j \in e} tf(a_j,c)} \cdot P(e|c) \right) \end{aligned} \tag{4.4}$$

Cette formule combine l'intérêt des deux premières formules. En effet, elle donne plus d'importance aux attributs fréquents dans la classe et moins à ceux qui appartiennent aux entités possédant beaucoup d'attributs.

Selon les trois sélections déterminées dans la section 4.2.1.1, la probabilité $P(e|c)$ est calculée selon l'appartenance de l'entité e à la sélection traitée E (*Shorty, Maxy* ou *Fantasy*). Dans notre étude, nous souhaitons que $P(e|c)$ pour $e \in E$ soit le double de $P(e|c)$ pour $e \notin E$ (ce choix est fait à des fins expérimentales, dans le seul but de favoriser les attributs des entités de la sélection traitée).

Après calcul, ces valeurs sont estimées comme suit :

$$\begin{cases} P(e|c) = 2/(|Inst(c)| + |Inst(E)|) & si\ e \in\ E \\ P(e|c) = 1/(|Inst(c)| + |Inst(E)|) & sinon \end{cases} \tag{4.5}$$

avec

— E : l'ensemble des entités à valoriser qui peut être *Shorty, Maxy* ou *Fantasy.*

— $|Inst(c)|$ et $|Inst(E)|$: le nombre d'entités dans la classe c et la sélection traitée E

Les probabilités vérifient que $\sum_{e \in c} P(e|c) = 1$.

Une fois que les données sont collectées et triées pour une requête classe, nous les agrégeons dans un tableau.

4.2.1.3 Évaluation

Afin d'évaluer notre approche, nous avons mené une évaluation utilisateur basée sur 30 requêtes, évaluées par 4 volontaires. 9 tableaux différents avec les 5 entités et les 10 attributs les plus pertinents ont été évalués (Formule 1 sans sélection, Formule 2 sans sélection/Shorty/Maxy/Fantasy, Formule 3 sans sélection/Shorty/Maxy/Fantasy). Les entités et les attributs ont tous été extraits de DBPedia, les attributs ont été filtrés afin de supprimer les informations répétitives, inutiles et de mise en forme. On trouvera les détails de notre expérimentation dans (Krichen et al., 2011, 2012). De manière synthétique, les conclusions de cette expérimentation sont les suivantes :

— les entités utilisées pour le calcul du poids de pertinence des attributs ont une importance. La sélection d'entités *Fantasy* (basée sur les liens entrants et sortants de l'entité dans Wikipedia) semble être la plus intéressante dans notre cas.

— le choix des fonctions utilisées pour pondérer les attributs impacte également significativement les résultats, la formule 3 étant la plus pertinente.

— La majorité des formules ont une MAP au-dessous de 0,5 même si P@10 permet d'obtenir relativement de bonnes performances. Une telle constatation nous conduit à conclure que les attributs pertinents sont bien parmi les dix premiers mais pas forcément bien placés dans les tableaux, ce qui nous pousse à améliorer davantage notre approche.

Les perspectives à court terme de ces travaux sont nombreuses, les cas des attributs multivalués ou évoluant dans le temps restant à traiter.

4.2.2 Construction de résumés temporels

Afin de construire un résumé temporel sur une entité, notre approche cherche à détecter en temps réel les phrases introduisant de nouvelles informations vitales (pertinentes et opportunes) à partir d'un flux de documents issus du Web. Par conséquent, elles doivent être pertinentes (concerner l'entité), exhaustives (couvrir les différentes informations publiées sur l'entité), non redondantes (reportées une seule fois) et émises sans trop de latence (Abbes et al., 2015a,b).

Formellement, considérons un flux continu F composé de documents d ayant chacun une date de publication $t(d)$ 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 $h_0, h_1, \ldots, h_m$ 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 $\forall d \in F_{h_i}, h_{i-1} \leq t(d) < h_i$.

Require: F : Flux de documents
Require: e : Entité à mettre à jour, ayant une étiquette $e.label$
Require: h_0 : Début de la période d'analyse du flux
Require: h_n : Fin de la période d'analyse du flux
Ensure: $V(e) \leftarrow \{\}$: L'historique des phrases vitales relatives à e

```
for i ∈ [1, n] do
    D_{h_i} ← sélection_des_documents(F_{h_i}, e.label)
    for d ∈ D_{h_i} do
        for s_j ∈ d do
            if contient_information_vitale(s_j, e) and est_nouvelle(s_j, V(e)) then
                enrichir(V(e), s_j)
            end if
        end for
    end for
end for
```

Algorithme 3 : Détection des phrases vitales relatives à une entité

L'algorithme 3 décrit le fonctionnement général de notre approche de détection des phrases vitales relatives à une entité donnée e. À chaque instant h_i, nous distinguons 3 étapes principales que nous détaillons dans les sous-sections suivantes :

1. la sélection des documents vitaux D_{h_i} par rapport à e en utilisant comme requête le ou les labels associés à l'entité dans la base de connaissances ;
2. la sélection des phrases vitales candidates (contenant une information vitale) ;

3. la vérification de la nouveauté des phrases candidates par rapport aux phrases appartenant à $V(e)$.

Nous considérons ici la sélection des documents vitaux faite selon une des approches du chapitre 3 et nous focalisons sur la détection des phrases vitales et de leur nouveauté.

4.2.2.1 Sélection des phrases vitales

Dans cette étape, nous analysons les phrases contenues dans les documents sélectionnés. Pour chaque phrase, nous devons décider si elle est vitale (reportant une information pertinente et opportune) par rapport à l'entité à surveiller e. Notre intuition est de considérer une phrase comme vitale si :

— elle est à proximité de l'entité e (des termes de l'étiquette de e),

— elle contient des mots « importants » relativement à l'entité e.

La proximité d'une phrase par rapport à l'entité e peut refléter sa pertinence. Une phrase mentionnant l'entité a plus de chance de parler de celle ci. Nous traduisons ainsi la proximité entre une phrase s_j et l'entité e en un score calculé selon l'équation suivante :

$$score\_proximité(s_j, e) = \frac{1}{|e.label|} \sum_{t \in e.label} \sum_{dist=0}^{dmax} e^{-dist * match\_s(t, s_{j+dist}, s_{j-dist})} \tag{4.6}$$

Avec

— $e.label$ est l'étiquette décrivant l'entité e.

— $|e.label|$ est le nombre de mots dans l'étiquette $e.label$.

— $match\_s(t, s_x, s_y)$ 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 considérons uniquement les phrases à proximité de l'entité e, c.à.d, ayant un *score proximité* supérieur à un seuil τ_p (la valeur de τ_p peut être déterminée expérimentalement).

En plus de la proximité, nous supposons que pour une entité e, il existe un ensemble de mots « importants » qui peuvent refléter la vitalité d'une phrase. Nous appelons ces mots des **mots déclencheurs**. Nous posons l'hypothèse selon laquelle les entités de même type (représenté dans la 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 `box-office`, `award`, `actrice`, `acteur`, `réalisateur` pourront être très utiles pour décrire les entités de type `film` car ils sont présents dans les annotations associées aux entités `Les aventures d'Alice au pays des Merveille` et `De l'autre côté du miroir`.

Formellement, soit $X(e) = \{A(e_1), A(e_2), \ldots, A(e_m)\}$ l'ensemble des m extraits des valeurs des annotations associées aux entités de même type que e. Nous pondérons les mots t par l'équation suivante :

$$\omega(t) = \frac{\sum_{i=1}^{m} TF(t, A_{e_i})}{IIF(t)} \tag{4.7}$$

— $TF(t, A_{e_i})$ est le nombre d'occurrences du terme t dans l'annotation A_{e_i},

— $IIF = log(\frac{m+1}{IF(t)})$ 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,

no	Date	Texte
s1	5 March 2010 - 07 :27	Alice in Wonderland has grossed $210,123,678 in North America
s2	25 March 2010 - 20 :50	Alice in Wonderland has grossed $334,191,110 in North America
s3	28 March 2010 - 06 :17	After two weeks of exploitation in U.S., Tim Burton's Alice made $334,191,110.

TABLEAU 4.1 – Exemple de phrases vitales.

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

Les ***top-k*** premiers mots seront considérés comme des mots déclencheurs pour l'entité e. Pour qu'une phrase soit considérée comme une phrase vitale candidate, il faut :

— que le score de proximité de la phrase soit $> \tau_p$,

— qu'elle contienne un mot déclencheur.

4.2.2.2 Détection de la nouveauté

Les phrases sélectionnées à l'étape précédente pourraient contenir des informations vitales redondantes déjà émises. Afin d'éliminer la redondance, nous comparons chaque phrase vitale candidate à toutes les phrases vitales déjà ajoutées à l'ensemble incrémental $V(e)$. Détecter la nouveauté n'est pas une tâche facile. Comme le montre le tableau 4.1, les deux phrases s1 et s2 contiennent un grand nombre de chaînes de caractères en commun, mais reportent deux informations différentes. Inversement, les phrases s2 et s3 ne sont pas trop similaires textuellement mais contribuent à la même information.

Dans notre méthode, nous proposons de prendre en compte les propriétés définies dans l'ontologie pour le concept type de l'entité que nous considérons. Nous cherchons à identifier dans la phrase, des entités ou des valeurs potentiellement liées sémantiquement car leur type correspond au domaine ou co-domaine des propriétés définies dans l'ontologie pour le concept.

Une phrase vitale candidate s_j est nouvelle si elle respecte la fonction de nouveauté suivante :

$$est\_nouvelle(s_j, V(e)) = texte\_divergent(s_j, V(e)) \circ entités\_liées(s_j, V(e)) \tag{4.8}$$

$$texte\_divergent(s_j, V(e)) = \begin{cases} faux & si\ \exists s_k \in V(e),\ cos(s_j, s_k) > \tau_n(V(e)) \\ vrai & sinon \end{cases} \tag{4.9}$$

$$entités\_liées(s_j, V(e)) = \begin{cases} vrai & si\ \exists x \in entités(s_j),\ \forall s_k \in V(e)\ x \notin entités(s_k) \\ faux & sinon \end{cases} \tag{4.10}$$

— Le symbole $\circ$ de l'équation 4.8 peut être un opérateur ***ET*** pour rendre le système orienté Précision (limiter la redondance), ou bien un opérateur ***OU*** pour privilégier le Rappel (l'exhaustivité)

— *entités(s_i)* est l'ensemble des entités liées reconnues dans la phrase s_i.

— $\tau_n(V)$ 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 :

$$\tau_n(V(e)) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{|V(e)|^2}{\delta^2}} \tag{4.11}$$

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

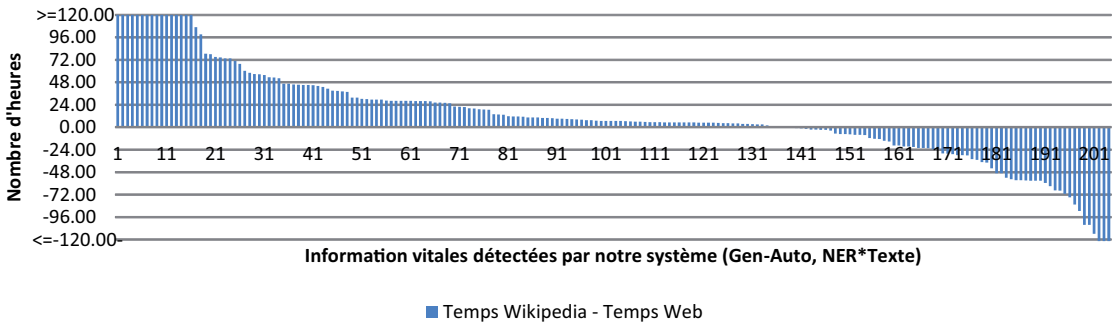


FIGURE 4.8 – Évaluation de la rapidité de notre système (Gen-Auto ; NER*Texte) par rapport aux mises à jours Wikipedia (Abbes et al., 2015a).

4.2.2.3 Évaluation

Notre approche a été évaluée sur les tâches TREC *Temporal Summarization* 2014 et 2015 (Aslam et al., 2013a, 2015). Ces tâches visent à construire des résumés temporels sur des entités de type événements (catastrophes naturelles, accidents...). Les résultats montrent l'importance de la combinaison conjonctive de la divergence textuelle (DIV) avec l'identification de nouvelles entités liées (NEL) pour la détection de nouveauté. Si l'on se compare à l'état de l'art, notre système aurait été classé 3^{e} sur la tâche 2014.

Notre approche pourrait également permettre la mise à jour temps-réel des bases de connaissances. La figure 4.8 compare la rapidité de notre système (dans sa meilleure configuration) à détecter les informations vitales pour les 24 événements de la tâche 2014 par rapport aux mises à jour Wikipedia. Notre système permet de détecter 67% d'informations vitales avant que celles-ci ne soient mises à jour dans Wikipedia. La moitié des informations sont détectées par notre système 7 heures avant qu'elles ne soient mises à jour dans Wikipedia. En moyenne, notre système permet de gagner 18 heures.

Dans le tableau 4.2, nous illustrons quelques exemples d'informations vitales détectées par notre système avant qu'elles soient ajoutées dans Wikipedia.

Id	*Information vitale détectée*	t_{web}	t_{wp}	t_{IB}	*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

TABLEAU 4.2 – Exemple d'informations vitales détectées par notre approche dans sa meilleure configuration. t_{web}, t_{wp}, t_{IB} représentent respectivement les temps de la disposition de l'information par notre système, dans Wikipedia et dans les infoboxes de Wikipedia. Extrait de (Abbes et al., 2015a).

La figure 4.8 et les exemples du tableau 4.2 montrent que les informations sont généralement publiées dans le documents Web (presse, blogs, etc.) avant qu'elles soient éditées dans les encyclopédies collaboratives comme Wikipedia. Notons que la mise à jour n'est pas forcément reportée dans les InfoBoxes principalement exploitées pour enrichir DBpedia. Bien que les entités analysées représentent des événements largement connus, qui intéressent plusieurs contributeurs, nous remarquons toujours un temps de latence, qui sera probablement plus grand pour des entités « moins connues ».

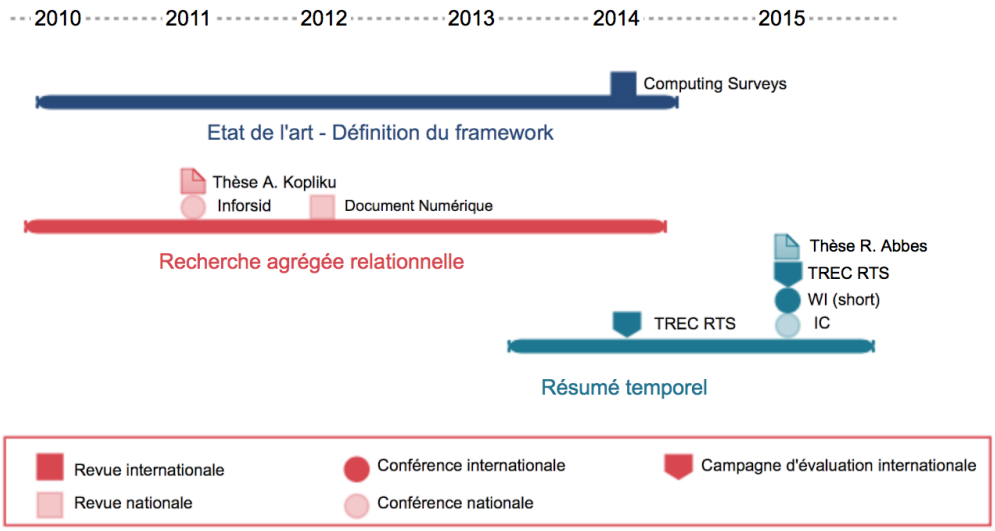


FIGURE 4.9 – Graphique synthétique de la structuration et de la valorisation des travaux sur le thème Agrégation autour des entités

4.3 Synthèse des travaux présentés

Nous avons proposé deux agrégats possibles pour deux problématiques de la recherche agrégée :

— la construction de tableaux pour des requêtes classe, comprenant des entités et attributs représentatifs. Ces tableaux sont construits à partir de relations entité-attributs donc l'extraction a été décrite au chapitre 3. Les opérations basiques d'agrégation utilisées ici sont le tri et la fusion [10].

— la génération de résumés temporels pour des requêtes entité. L'approche que nous proposons, comparée à l'état de l'art, nous permettrait de nous classer dans le top 5 des participations officielles à TREC *Temporal Summarization* 2014. Nos résumés temporels sont construits à partir de documents vitaux pour l'entité dont le filtrage a été présenté au chapitre 3, c'est-à-dire de documents apportant des informations nouvelles sur l'entité. Les opérations basiques d'agrégation sont cette fois le découpage, l'extraction (des phrases) et la fusion (pour former le résumé final).

Nos approches ont été évaluées soit par des études utilisateur (construction d'un tableau pour des requêtes classe), soit sur des collections d'évaluation de l'état de l'art (TREC *Temporal Summarization* 2014 et 2015). Nous avons également confronté ces approches à d'autres participants lors des campagnes officielles 2014 (Abbes et al., 2014a) et 2015 (Abbes et al., 2015c).

Diffusion scientifique. La figure 4.9 résume de façon chronologique mes différentes publications relatives aux sujets abordés dans le chapitre.

Formation à la recherche. J'ai co-encadré deux étudiants de Master sur ces thématiques (Damak, 2010; Krichen, 2010).

Les deux thèses mentionnées dans le chapitre précédent ont également des contributions largement mentionnées dans le chapitre :

— Arlind KOPLIKU (Kopliku, 2011), avec Mohand BOUGHANEM,

10. Pour rappel, ces opérations, décrites en section 4.1, sont le tri, le regroupement, la fusion (*merging*), le découpage et l'extraction.

— Rafik Abbes (Abbes, 2015), avec Mohand Boughanem et Nathalie Hernandez.

Projets. Une partie de ces recherches a été menée dans le cadre du projet PEPS RICA (*Recherche d'Information fondée sur le contexte et l'agrégation* - 2013-2014 [11]), puis du projet ANR CAIR (*Contextual Aggregated Information Retrieval* - 2015-2018 [12]). Nous nous sommes intéressés dans ces projets à l'élaboration de solutions pour les requêtes dites *agrégatives*, en nous focalisant sur les requêtes de type entité.

11. https://www.irit.fr/RICA/, dernier accès en mars 2018.
12. https://www.irit.fr/CAIR/fr/, dernier accès en mars 2018.

CHAPITRE 5

Recherche d'information et évaluation

> "And how many hours a day did you do lessons?" said Alice, in a hurry to change the subject.
> "Ten hours the first day," said the Mock Turtle: "nine the next, and so on."
> "What a curious plan!" exclaimed Alice.
> "That's the reason they're called lessons," the Gryphon remarked: "because they lessen from day to day."

Sommaire

Ce chapitre transversal aux précédents présente les recherches que nous avons menées dans le domaine de l'évaluation des SRI, et plus précisément la création de protocoles et collections de test.

5.1 Contexte des travaux

Le domaine de la RI a une forte tradition d'évaluation des performances. Ces dernières peuvent être évaluées d'un point de vue ingénierie et efficience (vitesse, coût d'accès/stockage), d'un point de vue utilisateur ou encore en fonction de leur efficacité (c'est-à-dire leur capacité à renvoyer des documents pertinents) (Saracevic, 1995). Nous nous sommes particulièrement intéressés, dans nos travaux, à l'évaluation automatique de l'efficacité des approches.

Depuis les années 1970, l'efficacité des SRI est principalement évaluée selon le paradigme de Cranfield (Cleverdon, 1967), reposant sur le principe des collections de test. Les collections de test sont composées des éléments suivants (Sanderson, 2010a) :

— un corpus de documents fixe,

— un ensemble de besoins d'information (souvent traduits sous forme de requêtes),

— des jugements de pertinence (ou vérité terrain), indiquant quels sont les documents pertinents pour chaque requête,

— des mesures d'évaluation permettant de mesurer quantitativement la performance des systèmes.

Fixer les contenus et les requêtes permet de comparer les systèmes entre eux, la reproductibilité des expérimentations, ainsi que la possibilité de comparer de nouvelles approches avec des résultats obtenus précédemment.

Les collections de test sont généralement créées dans le contexte de campagnes d'évaluation internationales et distribuées à la communauté dans ce cadre. Parmi les campagnes d'évaluation les plus connues, on peut citer TREC (*Text REtrieval Conference*, `http://trec.nist.gov`), CLEF (*Conference and Labs for the Evaluation Forum*, `http://www.clef-initiative.eu/`), NTCIR (*NII Test based Community for Information access Research*, `http://ntcir.nii.ac.jp/`) ou encore jusqu'à récemment INEX (*INitiative for the Evaluation of XML retrieval*, `http://inex.mmci.uni-saarland.de/`). Historiquement ces campagnes ont proposé d'évaluer la recherche *adhoc* (à partir d'une requête exprimée par l'utilisateur, le but est de restituer les seuls documents qui correspondent à son besoin), sur ces collections de documents variées. Ces dernières années, les tâches de recherche ont grandement évolué, en témoignent par exemple la tâche *Temporal Summarization* de TREC (Aslam et al., 2013b) où la vérité terrain est composée de termes, ou encore la tâche TREC *Open Search* dans laquelle l'évaluation se fait en direct via de vrais utilisateurs (Balog et al., 2014).

La grande majorité des travaux présentés dans ce mémoire a été évaluée dans le cadre des campagnes TREC, INEX ou encore CLEF (Abbes et al., 2013b, 2014a,b, 2015c; Ben Jabeur et al., 2012a, 2013; Chellal et al., 2015; Damak et al., 2011; Hlaoua et al., 2007e; Hubert et al., 2017b; Laitang et al., 2012a; Moulahi et al., 2016; Torjmen et al., 2007a, 2008c,d,e, 2009a,b). Cette connaissance des rouages de l'évaluation nous a valu d'être sollicités dans le cadre du projet européen Quaero [1], pour la mise en place de campagnes d'évaluation et de création de collections de test sur des domaines spécifiques :

— le *clustering* [2] de documents : le but de ces approches est de former des groupes de documents similaires (généralement par rapport à leur contenu) en réponse à une requête utilisateur (Manning et al., 2008),

1. Projet mené de 2008 à 2013, `http://www.quaero.org`, dernier accès mars 2018.
2. Dans la littérature française, plusieurs termes sont employés pour désigner le clustering : catégorisation, classification non supervisée ou classification automatique de documents. Le vocabulaire employé ne faisant pas consensus, nous préférons conserver le terme anglais dans ce mémoire.

— la recherche d'images par l'exemple : alors que l'expression du besoin en RI textuelle se fait généralement sous forme de mots-clés, en recherche d'images la requête peut également être une image exemple (`trouve moi toutes les images qui ressemblent à cette image, qui ont le même paysage, le même objet, ...`).

Construire des collections de test n'est pas chose aisée : il faut obtenir/collecter des documents, associer des requêtes « intéressantes », collecter les jugements de pertinence pour chaque requête et associer des mesures de pertinence appropriées pour la tâche de recherche considérée. Les sections 5.2.1 et 5.2.2 décrivent les différentes collections que nous avons créées.

Outre le montage de collections de tests, certains de mes travaux concernent plus particulièrement les protocoles d'évaluation : le protocole d'évaluation de la recherche agrégée inter-verticale (section 5.2.3), ainsi que le protocole d'évaluation d'une tâche TREC particulière, la tâche TREC *Real Time Summarization* (2016 et 2017). Ces derniers travaux ont donné lieu à la découverte de biais dans l'évaluation, biais que je détaille dans la section 5.2.4.

5.2 Contributions au domaine de recherche

5.2.1 Évaluation du clustering de documents

5.2.1.1 État de l'art de l'évaluation

L'approche classique des SRI est de renvoyer aux utilisateurs en réponse à une requête une liste de documents triés par pertinence système décroissante. Le clustering des résultats de recherche fait partie des alternatives proposées dans la littérature afin de faciliter l'accès à l'information. Le but des approches de clustering de résultats est de former des groupes de documents similaires en réponse à une requête utilisateur.

On trouvera dans (Carpineto et al., 2009) un panorama complet de ces approches de clustering. Elles sont présentes dans des systèmes aussi bien académiques que grand public tels que Vivisimo (Koshman et al., 2006) ou Exalead dans les années 2010 (`http://www.exalead.fr`, dernier accès mars 2018).

Historiquement, l'intérêt du clustering en RI a été souligné par Jardine et Van Rijsbergen (1971) à travers la définition de la *Cluster hypothesis*. Cette dernière stipule que l'appartenance de documents à un même groupe donne une indication quant à la pertinence de ces documents à une même requête. Autrement dit, les documents proches (concentrés dans un même groupe) tendront à être pertinents pour les mêmes requêtes. Dans le même temps, la *Cluster hypothesis* a suscité des critiques (Lamprier et al., 2010; Tombros, 2002). En effet, l'hypothèse qu'un système arrive à concentrer tous les documents pertinents au sein d'un même groupe est optimiste et lorsqu'un seul groupe est restitué, l'utilisateur doit tout de même le parcourir. Il se retrouve alors confronté aux mêmes écueils qu'avec une liste de résultats. L'utilisateur gagnerait peut-être à obtenir plusieurs groupes, dans lesquels les différents aspects de la requête seraient distribués.

Mesures. La validité des résultats renvoyés peut être évaluée de façon interne ou externe (Halkidi et al., 2001). Les mesures de validité interne vérifient certaines propriétés structurelles des groupes comme leur homogénéité, la détection des « individus » aberrants... (Carpineto et al., 2009). D'une manière générale, ces mesures cherchent à évaluer si le SRI atteint une forte similarité intra-cluster (les documents d'un même groupe sont similaires) et une faible similarité inter-clusters (les documents provenant de groupes différents sont différents) (Ingaramo et al., 2008).

La plupart des mesures de validité externe se basent sur l'existence d'une vérité terrain binaire (pertinent/non pertinent). Parmi elles on peut citer :

— celles cherchant à évaluer le groupe optimal (Jardine and van Rijsbergen, 1971; Tombros, 2002) ;

— celles cherchant à évaluer la qualité globale des groupes (Nayak et al., 2009; Vries et al., 2011) ;

— celles reconstruisant des listes de résultats par simulation d'usagers, pour ensuite effectuer une évaluation « traditionnelle » (Lamprier et al., 2010; Leuski, 2001).

Dans le premier cas (évaluation du groupe optimal), le meilleur groupe parmi l'ensemble des groupes est sélectionné en fonction de la précision, du rappel ou de la F-mesure. En sélectionnant le groupe optimal, il est ensuite possible de comparer les approches avec celles proposant des listes triées de résultats, en utilisant la mesure du Mk-1k (Jardine and van Rijsbergen, 1971; Tombros, 2002). Le principe est le suivant : soit l la liste des documents renvoyés par un moteur de recherche et utilisée en entrée de la technique de clustering évaluée. e_c et e_l sont deux ensembles de documents composés respectivement des documents du meilleur groupe de l'approche de clustering et des documents de l, avec l coupée à N documents, N étant le nombre de documents du meilleur groupe. Les ensembles e_c et e_l sont alors comparables grâce aux mesures rappel, précision et F_1 définies comme suit, pour un système s et une requête q :

$$Rappel(q) = \frac{RestPert(e_s, q)}{Pert(l, q)} \tag{5.1}$$

$$Precision(q) = \frac{RestPert(e_s, q)}{Rest(e_s, q)} \tag{5.2}$$

$$F_1(q) = \frac{2 \cdot Rappel(q) \cdot Precision(q)}{Rappel(q) + Precision(q)} \tag{5.3}$$

$Rest(e_s, q)$ et $RestPert(e_s, q)$ sont respectivement le nombre de documents restitués et le nombre de documents pertinents restitués par e_s pour q, avec $e_s \in \{e_c, e_l\}$. $Pert(l, q)$ est le nombre de documents pertinents dans l pour q. Un système idéal aura un rappel et une précision de 1. La mesure Mk-1k d'effectivité optimale est complétée par d'autres mesures, MK3 et MK4, dont on pourra trouver la description dans (Tombros, 2002).

Pour évaluer la qualité globale des groupes, il est possible d'utiliser la mesure nCCG (*Normalized Cumulative Cluster Gain*) (Nayak et al., 2009). Le gain cumulatif d'un groupe (CCG - *Cumulative Gain of a Cluster*) est calculé en comptant le nombre de documents pertinents du groupe. Pour une approche donnée, un vecteur trié CG est créé représentant chaque groupe en fonction de sa valeur CCG. Les groupes ne contenant pas de document pertinent ont une valeur de 0. Un vecteur de gain cumulé pour le vecteur CG est ensuite calculé et normalisé par le vecteur de gain idéal. Chaque solution de clustering cs est ensuite évaluée sur sa capacité à répartir l'information pertinente sur les groupes pour une requête q. On a ainsi :

$$SplitScore(q, cs) = \sum^{|CG|} \frac{cumsum(CG)}{Pert(q)^2} \tag{5.4}$$

avec $Pert(q)$ le nombre de documents pertinents dans l'ensemble total de résultats pour la requête q. Le pire scénario est celui dans lequel chaque document pertinent est dans un groupe différent. Soit CG1 le vecteur qui contient le gain cumulatif de chaque groupe.

$$MinSplitScore(q, cs) = \sum^{|CG|} \frac{cumsum(CG1)}{Pert(q)^2} \tag{5.5}$$

Le gain cumulatif normalisé est alors :

$$nCCG(q, cs) = \frac{SplitScore(q, cs) - MinSplitScore(q, cs)}{1 - MinSplitScore(q, cs)} \tag{5.6}$$

La mesure nCCG est comprise entre 0 et 1 : une bonne solution de clustering a une valeur élevée, ce qui traduit le fait qu'un grand nombre de documents pertinents sont regroupés dans un même groupe.

Les mesures précédentes se basent sur le fait que les documents répondant à une même requête sont dans un même groupe (respect de la *cluster hypothesis*). Or, lorsqu'une requête revêt plusieurs aspects, les groupes peuvent être vus comme complémentaires, et donc ces mesures ne sont plus adaptées. Pour solutionner le problème, plusieurs approches permettent de simuler le comportement de l'usager dans le parcours des groupes. Cette simulation, basée sur le fait que les groupes sont triés, ainsi que les documents qu'ils contiennent, aboutit à la reconstruction d'une liste de résultats, qui sera ensuite évaluée selon des méthodes « traditionnelles ». Les parcours les plus simples sont les parcours en profondeur et en largeur. Le parcours en profondeur examine les groupes les uns après les autres, en considérant tous les documents d'un groupe avant de passer au suivant. Le parcours en largeur considère le premier document non lu par groupe, en bouclant sur l'ensemble des groupes, jusqu'à ce que tous les documents aient été lus. Leuski (2001) propose de simuler le comportement de l'utilisateur qui explorerait un groupe séquentiellement et en changerait dès qu'il trouverait plus de documents non pertinents que de pertinents. Les travaux de (Lamprier, 2008; Lamprier et al., 2010) proposent également plusieurs parcours : par exemple, le parcours orienté par la pertinence des documents modélise un usager qui prend en compte le ratio de documents pertinents trouvés dans chaque groupe, et le parcours orienté par la proximité des documents pertinents modélise un usager qui oriente sa recherche vers des groupes dont le contenu des documents examinés semble correspondre aux informations portées par les documents pertinents.

Dès lors que les jugements de pertinence reflètent la variété des résultats (c'est-à-dire dès lors que la pertinence n'est plus binaire mais montre les différents aspects de la requête), ou que l'on connaît *a priori* les catégories des documents, les groupes résultats peuvent être comparés à la classification idéale (*Gold Classification*). La mesure la plus utilisée dans ce cadre est probablement la pureté. La pureté mesure à quel point un groupe contient des documents majoritairement d'une seule classe/catégorie. Chaque groupe c se voit assigner le label correspondant à la majorité des documents qu'il possède.

$$\text{Pureté}(c) = \frac{\text{nombre de documents avec le label majoritaire dans c}}{\text{nombre de documents dans c}} \tag{5.7}$$

Puisqu'il y a plusieurs documents renvoyés pour chaque requête (pour une solution de clustering cs), les valeurs de pureté peuvent être agrégées selon une micro ou macromoyenne.

$$\text{micropureté}(cs) = \frac{\sum_{k=0}^{n} \text{pureté(k)} \cdot \text{nombre de documents dans k}}{\text{nombre de documents renvoyés dans cs}} \tag{5.8}$$

$$\text{macropureté}(cs) = \frac{\sum_{k=0}^{n} \text{pureté(k)}}{n} \tag{5.9}$$

Dans un contexte de recherche d'information, il est très facile d'obtenir une micro et une macro-pureté élevées, il suffit de renvoyer autant de groupes que de documents (c'est-à-dire un seul document par groupe). Pour limiter ce biais, le nombre de groupes peut être comparé au nombre de catégories de la classification idéale, ou on peut encore considérer le nombre de catégories qui ont été correctement identifiées.

Parmi les autres mesures servant à comparer avec une classification idéale, on peut citer la mesure F1, ou encore l'entropie et l'information mutuelle (Crabtree et al., 2005). Ces mesures sont cependant seulement informatives puisqu'elles ne permettent pas vraiment de comparer deux méthodes entre elles. De plus, elles dépendent beaucoup du nombre de groupes renvoyés par les différents modèles (nombre qui n'est pas forcément identique). Enfin toutes ces mesures ne peuvent pas être utilisées lorsque les documents peuvent faire partie de plusieurs catégories. Par exemple, la mesure d'information mutuelle nécessite qu'il n'y ait pas d'intersection entre les groupes (Crabtree et al., 2005).

Campagnes d'évaluation et collections de test A notre connaissance, il existe peu de campagnes d'évaluation permettant d'évaluer des techniques de clustering des résultats de recherche. La plupart des évaluations existantes, parmi lesquelles on peut citer celles se basant sur la collection Reuters (Lewis et al., 2004), évaluent la capacité des systèmes à faire des groupes de documents correspondant à des classes pré-établies mais indépendamment de toute requête.

Si l'on s'intéresse précisément à l'évaluation du clustering des résultats, nous pouvons citer la tâche *XML Mining* de la campagne d'évaluation INEX qui en 2009 et 2010 a explicitement concerné l'évaluation de la *Cluster hypothesis*. Elle s'est basée sur les requêtes et jugements de pertinence utilisés dans le cadre des tâches de recherche *adhoc* proposées les mêmes années. Il s'agissait d'évaluer la qualité des groupes dans le but de sélectionner le groupe de documents optimal pour chaque requête. La mesure NCCG présentée précédemment a été utilisée pour comparer les systèmes.

5.2.1.2 Collections de test créées

Dans le cadre du projet Quaero, nous avons créé deux collections de test, associées chacune à une campagne d'évaluation pour les participants au projet.

Notre première collection visait à évaluer la *Cluster hypothesis*, comme dans le cadre des campagnes INEX 2009 et 2010. Nous avons utilisé une collection composée de 2,6 millions de pages web issues du domaine français (.fr) et aspirées par le moteur de recherche Exalead en 2008. Nous avons associé à cette collection 25 besoins, comprenant une requête sous forme de mots-clés et un texte explicitant le besoin en information. Ces besoins (que nous nommons mono-aspect) sont des besoins réels qui ont été soumis par des utilisateurs d'Exalead (extraits du *log* de ce moteur de recherche). Les jugements de pertinence ont été recueillis suivant une procédure similaire à celle utilisée dans TREC. Nous avons constitué un *pool* de résultats issus de 144 configurations différentes du moteur de recherche Terrier (Ounis et al., 2005). Ces configurations ont été construites en utilisant différentes formes d'indexation, différents modèles de recherche, et en réalisant ou non l'expansion de requêtes. Ce *pool* a ensuite été évalué manuellement pour identifier les documents pertinents de chaque requête (pertinence binaire : pertinent/non pertinent). L'outil utilisé a été développé dans notre équipe, et nous avons été en charge du recrutement des assesseurs. L'évaluation a ensuite été faite selon la mesure Mk-1k décrite précédemment.

Cette première collection a notamment été utilisée pour évaluer le système de clustering Kodex, basé sur la détection de communautés sur le graphe biparti documents-termes (Navarro, 2013; Navarro et al., 2011).

La seconde collection avait un double objectif. Le premier était de nouveau d'évaluer des approches suivant la *Cluster hypothesis*, avec des requêtes « mono-aspect ». Le second était d'évaluer des approches remettant en cause la *Cluster hypothesis*, et formant des groupes représentatifs des différents aspects d'une requête. Pour ce faire, nous avons utilisé un corpus de 10 millions de documents collectés par Exalead sur une période de 3 mois (100 premiers résultats de toutes les requêtes sur le moteur pendant 3 mois). À l'aide des logs associés, nous avons sélectionné 25 requêtes mono-aspect comme pour la première campagne d'évaluation, mais également 25 requêtes multi-aspects pour lesquelles la pertinence pouvait être définie selon plusieurs aspects (par exemple, la requête `bol d'or` peut concerner une course de moto ou bien une course de voile). Ces aspects, déterminés *a priori* en fonction de documents du corpus, n'ont pas été fournis aux participants de la campagne, mais ont servi pour les jugements de pertinence. Là encore, pour former le *pool* de documents utilisé pour les jugements de pertinence, nous avons utilisé plusieurs configurations du moteur de recherche Terrier (112). Les jugements de pertinence pour les requêtes mono-aspect ont été effectués, comme pour la première campagne, de façon binaire. Pour les requêtes multi-aspects, les personnes effectuant les jugements de pertinence ont dû déterminer pour chaque document du *pool* l'aspect principal du document ainsi éventuellement qu'un ou plusieurs aspects se-

condaires. L'évaluation a ensuite été faite en fonction de la classification idéale (calcul de la pureté des groupes). À notre connaissance, cette dernière collection est la seule collection de RI permettant d'évaluer des techniques de clustering des résultats de recherche ne respectant pas la *Cluster hypothesis*.

5.2.2 Évaluation de la recherche d'images par l'exemple

5.2.2.1 État de l'art de l'évaluation

Nous nous intéressons principalement dans cette section aux problématiques d'évaluation lorsque la requête utilisateur est une image (éventuellement complétée par des mots-clés). Deux tâches de recherche ressortent dans ce contexte :

— la recherche *adhoc* (trouve les images similaires à cette image) ;

— la détection de copies (cette image est-elle la copie d'une autre image ? Y a-t-il des copies de mon image dans la collection ?). Cette tâche est cruciale dans un contexte de protection des droits d'auteur.

Les SRI capables de traiter ce genre de requêtes utilisent des approches de recherche d'images par le contenu (*Content-based Information Retrieval*) : il s'agit d'utiliser des caractéristiques visuelles des images telles que la couleur, la texture, les formes... D'une manière générale, les approches proposées calculent des descripteurs sur des caractéristiques visuelles globales (concernant l'image entière) et locales (concernant les objets de l'image) (Datta et al., 2005).

Mesures. En recherche d'images, l'efficacité et l'efficience sont très étroitement liées, de façon peut-être plus importante qu'en RI textuelle. En effet, le calcul des descripteurs des images peut être très coûteux (de nombreuses approches cherchant à réduire leur dimension (Datta et al., 2005)), il faut alors trouver le bon équilibre entre qualité des résultats et temps d'exécution. C'est pour cette raison que souvent les mesures d'efficacité sont corrélées avec des mesures d'efficience (Moëllic and Fluhr, 2006). Parmi les mesures d'efficience pouvant être prises en compte, on peut citer le temps de calcul des descripteurs, le nombre d'accès disque, le temps moyen de traitement d'une requête, la complexité de l'approche...

Les mesures utilisées dans le cadre d'une recherche *adhoc* sont très similaires à celles de la recherche *adhoc* textuelle : on peut citer des mesures telles que la MAP, BPref, R-Précision, P@r, nDCG... (Müller et al., 2001; Sanderson, 2010b).

Campagnes d'évaluation et collections de test. La campagne d'évaluation la plus connue en recherche d'images est ImageCLEF. ImageCLEF fonctionne sur le même principe que la campagne TREC, avec des tâches de recherche évaluées variant d'une année sur l'autre. ImageCLEF a été lancée pour la première fois en 2003, dans le cadre de la campagne d'évaluation CLEF (*Cross-Language Evaluation Forum*). Parmi les tâches d'évaluation partant d'une image requête, on peut citer :

— la tâche de recherche d'images photographiques (2003-2009) (Paramita and Grubinger, 2010), dans laquelle les requêtes sont constituées d'images exemple et d'un descriptif textuel ;

— la tâche de recherche d'images dans Wikipedia (2008-2011) (Tsikrika and Kludas, 2010), pour laquelle là encore les requêtes sont constituées d'images exemple et d'un descriptif textuel. Le but ici est de se confronter à des collections d'images plus grandes (230 000 images pour la collection Wikipedia) ;

— la tâche de recherche d'images médicales (2004-2012) (Müller and Kalpathy-Cramer, 2010) : là encore les requêtes sont constituées d'images exemple et d'un descriptif, mais les problématiques évaluées sont propres à la recherche médicale (recherche d'images

suivant une certaine modalité (radios, scanners, images provenant de microscopes...), recherche de cas cliniques...) ;

— la tâche de recherche d'objets ou de concepts (2007-2009) (Nowak et al., 2010) : il s'agit de rechercher des objets (voiture par exemple) ou des concepts (paysage, ...) au sein d'une collection d'images. La requête peut être fournie sous forme d'images exemple.

Une autre initiative a également vu le jour en 2006 en France : la campagne ImagEval (Moëllic and Fluhr, 2006), fondée par le programme français « Techno-Vision ». La campagne s'est intéressée à la recherche *adhoc*, la détection d'objets mais aussi la détection d'images transformées. Faute de financements, la campagne n'a malheureusement pas été poursuivie les années suivantes.

5.2.2.2 Collections de test créées

Toutes les campagnes précédemment citées travaillent sur des collections d'images de taille relativement petite. Nous avons participé, en partenariat avec Exalead et l'équipe TeXMeX de l'INRIA de Rennes, à la construction de collections de test de plus grande ampleur, utilisées pour des tâches de détection de copies. Exalead et l'équipe TeXMex ont fourni les données :

— une collection de 1 128 000 images en deux résolutions (512 et 150 pixels sur leur plus grande largeur). Les images de cette collection proviennent principalement de photos de Flickr auxquelles ont été ajoutées des collections utilisées traditionnellement en recherche d'images (Caltech (Li et al., 2004), INRIA holidays (Jégou et al., 2008)). 1 000 images sélectionnées aléatoirement ont subi 49 transformations différentes (rotation, compression (qualité JPEG), dégradation par ajout de bruit ou de textes...) afin de servir de requêtes (Petitcolas et al., 1998) ;

— une collection de 100 million d'images (98 064 203 exactement) collectées sur le web par Exalead. Ces images font sur leur plus grand côté 150 pixels, et le volume occupé par la collection est de plus de 2 To. Les requêtes utilisées sont les mêmes que pour la collection précédente.

Nous avons créé deux scénarios d'évaluation, dans un contexte de protection des droits d'auteur :

— S1 : Étant donnée une copie, trouver l'image originale dans la collection si elle existe (en d'autre termes : cette image semble être une copie, est-ce le cas ?),

— S2 : Étant donnée une image originale, trouver ses copies dans la collection (l'auteur d'une image veut savoir si elle a été utilisée dans d'autres contextes)

Nous avons proposé d'évaluer les approches en fonction de leur efficience (temps d'indexation, RAM nécessaire, taille de l'index, nombre d'accès disque, nombre de cœurs utilisés, complexité de chaque approche) et de leur efficacité. Pour évaluer l'efficacité du scénario S2, les mesures usuelles basées sur le rappel et la précision peuvent être utilisées (MAP - *Mean Average Precision), Rappel@X et Precision@X.*

Le scénario S1 est un peu différent : il se peut qu'il n'y ait pas d'image réponse dans la collection, et s'il y en a, il n'y en a qu'une. Nous avons donc proposé d'utiliser la MRR (*Mean Reciprocal Rank*) (Voorhees, 1999), c'est-à-dire le rang de la première bonne réponse d'un système, ainsi que l'Averaged MAP : pour chaque *run*, tous les résultats, y compris les résultats des requêtes pour lesquelles il n'y a pas de correspondant dans la collection sont concaténés et triés en fonction de leur score (ce qui implique que les scores sont comparables d'une requête à l'autre). La précision est ensuite évaluée à chaque point de rappel (c'est-à-dire à chaque image pertinente retrouvée). Nous avons ainsi :

$$AveragedMAP = \frac{1}{|Pert|} \sum_{j=1}^{|Pert|} Precision(j) \tag{5.10}$$

où $|Pert|$ est le nombre total d'images pertinentes dans la collection. Cette mesure permet ainsi de tenir compte des requêtes sans image réponse dans la collection.

Nous avons utilisé la première collection pour une campagne d'évaluation menée en interne dans le projet Quaero (avec 3 participants). La deuxième collection a quant à elle été utilisée à des fins de vérification de passage à l'échelle des algorithmes (Moise et al., 2013a,b). Elle est à ce jour et à notre connaissance une des plus grandes collections utilisée pour l'évaluation de la recherche d'images.

5.2.3 Évaluation de la recherche agrégée inter-verticale

5.2.3.1 État de l'art

Les termes « recherche agrégée inter-verticale » ont fait leur apparition à la fin des années 2000 (Lalmas, 2011). Les moteurs de recherche du Web mettent actuellement tous en place ce type de recherche agrégée, en incluant dans les résultats de recherche des plans, des vidéos, des images, ou bien encore des news. Lorsque nous nous sommes intéressés au problème en 2010, les recherches dans le domaine prenaient trois directions principales (Kopliku et al., 2011c) :

— une première direction concernait les interfaces de visualisation des résultats (Sushmita et al., 2009, 2010), en se demandant principalement s'il fallait réserver des encarts fixes aux verticaux dans les résultats de recherche. Il est aujourd'hui communément admis que les verticaux soient intégrés à la liste triée de résultats (Turpin et al., 2016), comme le fait par exemple Google. Les recherches actuelles dans le domaine s'intéressent maintenant à l'emplacement spécifique des verticaux dans la liste triée, ainsi qu'au nombre de résultats par vertical à présenter aux utilisateurs (Arguello and Capra, 2016) ;

— une seconde direction cherchait à prédire si les résultats de certaines sources (verticaux) devaient être inclus ou non dans les résultats de certaines requêtes (Arguello et al., 2009; Diaz, 2009; Li et al., 2008) ;

— une dernière direction s'intéressait à l'évaluation de l'intérêt de la recherche agrégée inter-verticale ainsi qu'aux nouveaux protocoles d'évaluation nécessaires.

L'**intérêt** de la recherche agrégée inter-verticale a été étudié dans (Arguello et al., 2009; Li et al., 2008). De premières recherches ont montré que l'intention du vertical est souvent présent dans les requêtes. Sushmita et al. (2009, 2010) ont quant à eux montré que la recherche agrégée inter-verticale augmente le nombre et la diversité des résultats pertinents proposés à l'utilisateur. Ces conditions, certes nécessaires, ne suffisent pas à prouver son intérêt. Si les résultats proposés par un vertical sont non pertinents ou que les résultats de deux verticaux sont redondants, la recherche agrégée inter-verticale perd de son sens.

D'autres recherches se sont intéressées à **évaluer l'efficacité** des systèmes. Certaines se sont concentrées sur l'évaluation de la sélection des sources (Arguello et al., 2009; Li et al., 2008; Liu et al., 2009), d'autres ont comparé les interfaces de restitution des résultats proposées (Sushmita et al., 2009, 2010; Thomas et al., 2010). Des jugements de pertinence humains (Arguello et al., 2009; Sushmita et al., 2009) ou bien basés sur des clics ont été considérés (Diaz, 2009; Sushmita et al., 2010).

Chacune de ces évaluations n'a considéré qu'un seul type de pertinence (par intention ou par le contenu réel), et qu'un seul type de requête (avec un besoin clairement identifié ou pas). Les résultats pourraient cependant ne pas être généralisables. Par exemple, la requête `photo de Johnny Depp` sans besoin clairement explicité et évaluée par intention, c'est-à-dire en ne considérant que la pertinence a priori des verticaux, laisserait croire que seul le vertical « Image » est pertinent. La visualisation de contenus de news ou de vidéos, et/ou la définition précise du besoin via un texte explicatif (l'équivalent d'un *narratif* dans le jargon TREC) pourrait cependant permettre d'inclure à la vérité terrain d'autres verticaux pertinents et donc d'autres granules pertinents pour la requête.

Afin de confirmer ou infirmer ces résultats, nous avons pour notre part construit une étude d'usage ayant pour but (Kopliku et al., 2011c) :

— d'évaluer l'intérêt de la recherche agrégée inter-verticale selon deux types de pertinence (par intention ou par le contenu) et deux types de requêtes (avec un besoin clairement identifié ou pas),

— d'évaluer l'impact de la pertinence et du type de requête sur l'évaluation.

5.2.3.2 Protocole expérimental

Nous définissons la pertinence de la façon suivante :

— **Pertinence par intention** : une source est pertinente pour une requête si cela a du sens de soumettre cette requête à ce vertical. La pertinence de la source est indépendante de la pertinence des résultats qu'elle pourrait renvoyer.

— **Pertinence par le contenu** : une source est pertinente pour une requête si elle contient des résultats pertinents pour la requête. La pertinence de la source est donc dépendante des résultats qu'elle renvoie.

Nous avons sélectionné aléatoirement 100 requêtes de la tâche TREC Million Query Track (Carterette et al., 2008). Ces requêtes sont toutes extraites de *logs* de moteurs de recherche. À chacune d'entre elles, nous avons associé un descriptif (l'équivalent du *narratif* des requêtes TREC), nous permettant ainsi pour chaque requête initiale d'avoir 2 formulations : quelques mots-clés (besoin non clairement identifié, nous appelons ces requêtes des **requêtes courtes**) ou bien une description plus complète (besoin clairement identifié, nous appelons ces requêtes **requêtes avec besoin fixe**). Nous avons considéré 9 sources (verticaux), issues de l'état de l'art : le Web, ainsi que la recherche de vidéos, d'images, de plans/cartes (recherche géographique), Wikipedia, la recherche de produits, le question-réponse et les définitions.

Nous avons évalué 4 configurations de recherche différentes :

1. requête courte, pertinence par intention ;
2. requête courte, pertinence par contenu ;
3. requête avec besoin fixe, pertinence par intention ;
4. requête avec besoin fixe, pertinence par contenu.

Afin d'évaluer la pertinence par le contenu, nous avons construit une interface d'évaluation (voir Figure 5.1). Les résultats sont affichés dans 9 panneaux formant un carré de 3*3. À chaque panneau est clairement associée une source (l'intitulé de la source est indiqué au-dessus du panneau). Il contient seulement les résultats de cette source. Afin d'éviter tout biais, les sources sont affichées dans un ordre aléatoire en fonction des requêtes. Le nombre de résultats par source est choisi en fonction de l'espace de visualisation : 9 images, 1 carte/plan et 3 jusqu'à 3 résultats pour les autres sources.

L'étude d'usage a été menée avec 33 participants de l'IRIT. Pour les tâches de recherche 1, 3 et 4, 10 participants ont évalué chacun 30 requêtes. La tâche de recherche 2, plus chronophage que les autres, a demandé la participation de 30 personnes pour 10 requêtes chacun. Les requêtes ont été réparties pour que chacune d'entre elles soit évaluée par 3 participants.

5.2.3.3 Conclusions

Les résultats détaillés de notre étude d'usage sont présentés dans (Kopliku et al., 2011c). Nous pouvons en retenir les points suivants :

— Concernant l'intérêt de la recherche agrégée inter-verticale :
  — la pertinence est distribuée sur les différents verticaux,
  — les verticaux peuvent répondre à de nombreuses requêtes. Ils sont souvent pertinents en même temps que la recherche Web, mais peuvent aussi parfois présenter des résultats que l'on ne retrouve pas sur une simple recherche Web ;

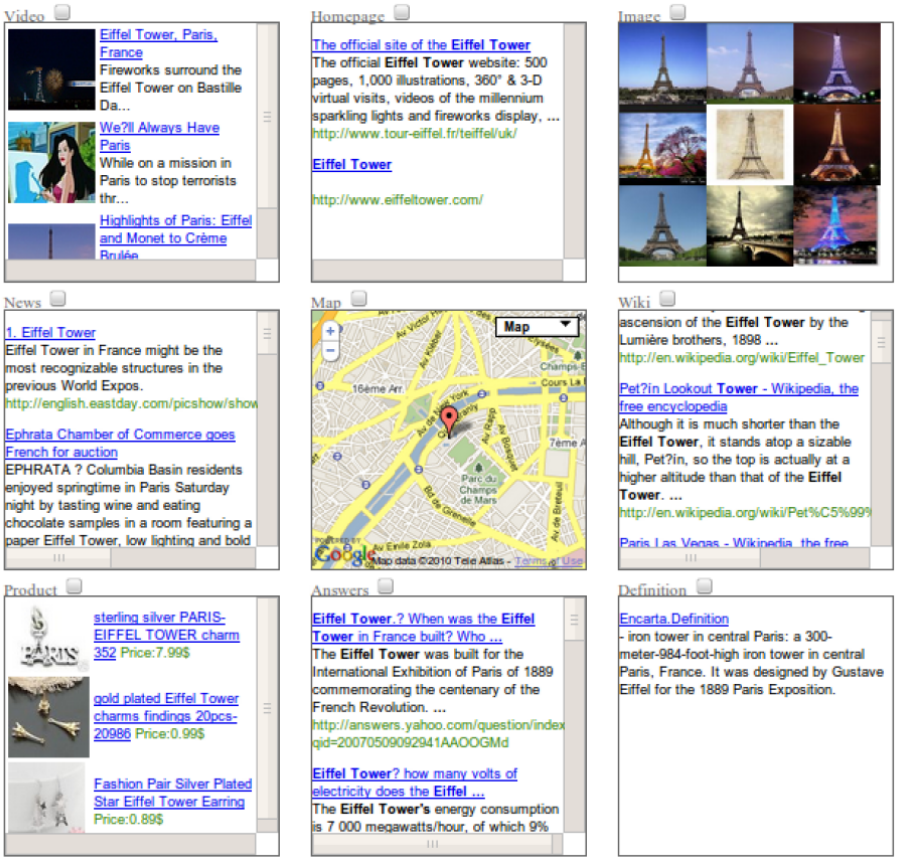


FIGURE 5.1 – Interface d'évaluation pour la recherche agrégée inter-verticale, image extraite de (Kopliku et al., 2011c).

— la plupart des requêtes possèdent plus d'un vertical pertinent.

Cela montre que la recherche agrégée inter-verticale est utile pour les utilisateurs, car une source est souvent insuffisante pour répondre au besoin. Les sources multiples peuvent se complémenter pour répondre de façon plus exhaustive.

— Concernant le protocole d'évaluation de la recherche agrégée inter-verticale :

— Clairement, considérer les requêtes sans descriptif n'est pas suffisant. Se contenter de requêtes issues de logs de moteurs de recherche sans leur associer un besoin peut donc générer des résultats biaisés. Par exemple, considérons la requête `Hamilton county`. Avec une requête courte, les sources « Images » et « Wikipedia » peuvent être pertinentes. Ce ne sera plus le cas si le besoin est explicité et décrit de la façon suivante `I want the location of Hamilton county`. Fixer le besoin en information diminue le nombre et la diversité des verticaux pertinents, mais facilite la tâche d'évaluation. Le temps d'évaluation est inférieur et l'accord inter-assesseurs supérieur.

— Les participants peuvent identifier plus de sources pertinentes par le contenu que par intention.

Se contenter d'évaluer la recherche agrégée inter-verticale avec des requêtes courtes (de *logs* de moteurs de recherche) et une pertinence par intention, est certes plus rapide, mais beaucoup moins réaliste, et peut conduire à des interprétations erronées. L'évaluation la plus rigoureuse et réaliste semble donc celle utilisant des requêtes avec besoin fixe et pertinence par contenu.

5.2.4 Protocole d'évaluation de la campagne TREC Real Time Summarization 2016 et 2017

5.2.4.1 Présentation de la tâche

Comme décrit au chapitre 2, une partie de mes recherches concerne la recherche temps-réel dans des microblogs. Notre approche se base sur des filtres successifs permettant de conserver les microblogs correspondant à des profils utilisateurs donnés. La tâche *Real Time Summarization* de la campagne d'évaluation TREC fournit un cadre d'évaluation idéal pour cette approche. Cette tâche, créée en 2016 à partir des tâches TREC *Microblog* et TREC *Temporal Summarization*, définit deux scénarios d'évaluation : le scénario A *Push notifications* qui correspond à un système envoyant immédiatement les microblogs considérés comme pertinents et le scénario B *Email digest* dans lequel les systèmes envoient une fois par jour un résumé des microblogs pertinents du jour (Lin et al., 2016, 2017). Nous nous sommes plus particulièrement intéressés au scénario A.
Le protocole d'évaluation est le suivant : chaque groupe participant doit traiter le flux Twitter publiquement accessible, correspondant à 1 % du nombre total de tweets publiés. La période d'évaluation est partitionnée en jours ; elle a duré respectivement 10 et 8 jours en 2016 et 2017. Un ensemble de profils (*topics*) est fourni aux participants, pour lesquels chaque système ne doit pas renvoyer plus de 10 tweets par jour. Lorsqu'aucun tweet pertinent pour le profil n'est publié un jour donné, les systèmes doivent rester silencieux et ne pas renvoyer de tweets, sous peine d'être pénalisés.

Deux évaluations différentes sont considérées : des jugements temps réels (*online judgments*) ayant lieu pendant la période d'évaluation et des jugements a posteriori (*batch judgments*). Les travaux étudiant ces deux modes d'évaluation ont montré qu'ils étaient corrélés (Roegiest et al., 2017; Tan et al., 2016). L'évaluation *online* n'étant pas réutilisable pour des travaux postérieurs à la campagne (Tan et al., 2017), nous nous sommes focalisés sur l'évaluation *batch*. Les jugements de pertinence *batch* sont faits sur un *pool* de tweets issus des participations officielles aux scénarios A et B. Les jugements sont faits en deux phases :

— dans un premier temps, les tweets se voient assigner un jugement de pertinence simple (non pertinent, pertinent, très pertinent)

— dans un second temps, les tweets sont répartis en clusters partageant la même information sémantique, via une approche nommée TTG (*Tweet Timeline Generation*) dans la littérature (Wang et al., 2015b). Les clusters sont ensuite considérés comme d'importance égale dans le protocole d'évaluation.

La notion de **pertinence** finale d'un tweet est fondamentale : un tweet est considéré comme pertinent **si et seulement si** il fait partie d'un cluster **et** qu'aucun autre tweet du cluster n'a déjà été renvoyé (dans ce dernier cas le tweet est considéré comme redondant).

5.2.4.2 Mesures d'évaluation

Dans les objectifs de la tâche, il est clairement indiqué que les systèmes doivent être efficaces (renvoyer des tweets de bonne qualité) et efficients (pas de latence dans le renvoi des tweets). Les organisateurs ont décidé de séparer les deux types d'évaluation, certains systèmes préférant privilégier l'efficience alors que d'autres privilégient l'efficacité.

Mesures d'évaluation orientées gain. Trois mesures basées sur le concept de gain sont utilisées pour évaluer la qualité des tweets renvoyés. Pour ces mesures, le gain est calculé comme suit : étant donné un jour w_j et $T_i(w_j)$ l'ensemble des tweets renvoyés par le système S_i et publiés durant w_j, le gain $G(w_j, S_i)$ est évalué comme suit :

$$G(S_i, w_j) = \sum_{t \in T_i(w_j)} g(t) \tag{5.11}$$

où $g(t)$ est le gain du tweet t : $g(t) = 1$ si le tweet est pertinent, $g(t) = 0$ sinon, c'est-à-dire que t est non pertinent ou redondant.

La mesure *EG* (*Expected Gain*) est adaptée de (Aslam et al., 2015). Étant donné un jour w_j, nous avons :

$$EG(w_j, S_i) = \frac{1}{|T_i(w_j)|} \cdot G(S_i, w_j) \tag{5.12}$$

où $|T_i(w_j)|$ est le nombre de tweets envoyés par S_i et publiés durant w_j.

Une question très importante sur cette mesure est la façon dont les jours silencieux sont considérés. Trois variantes de la mesure ont été introduites par les organisateurs de la tâche :

— *EG*-0 pour laquelle les systèmes reçoivent un gain de 0 pour les jours silencieux indépendamment des tweets qu'ils ont renvoyés ;

— *EG*-1 pour laquelle les systèmes reçoivent un gain de 1 pour les jours silencieux s'ils n'ont effectivement pas renvoyé de tweets publiés pendant la journée, et 0 sinon.

— *EG*-p pour laquelle la proportion de tweets retournés durant le jour silencieux est considérée : un système reçoit un score de $\frac{N-|\bar{t}|}{N}$, où $|\bar{t}|$ est le nombre de tweets non pertinents renvoyés par le système. Par exemple, si un système renvoie un tweet publié pendant le jour mais non pertinent (au lieu de 0), il reçoit un score de 0.9 ; 2 tweets non pertinents font baisser le score à 0.8, etc.

La façon dont les jours silencieux sont considérés est donc cruciale, puisqu'un impact énorme des jours silencieux sur les jours non silencieux a été observé dans l'évaluation (Tan et al., 2016).

Deux autres mesures détaillées dans (Lin et al., 2016, 2017) ont également été proposées : *nCG* (*Normalized Cumulative Gain*) et GMP (*Gain Minus Pain)*. Nos analyses sur ces mesures sont détaillées dans (Hubert et al., 2017a,b), nous ne rapportons ici que celles concernant la mesure *EG*, considérée comme mesure principale de la tâche.

Latence. La latence est définie comme :

$$Latency(S_k) = \sum_{t_i^{(\cdot)} \in R_k} \Pi_k(t_i^{(\cdot)}) - \Theta(t_i^1) \tag{5.13}$$

où t_i^1 est le premier tweet du cluster C_i, $t_i^{(\cdot)}$ est le tweet le plus ancien renvoyé par le système S_k pour le cluster C_i, et R_k est l'ensemble des tweets pertinents renvoyés par S_k. Les fonctions $\Pi_k(t)$ et $\Theta(t)$ permettent respectivement de connaître la date de renvoi du tweet et son étiquette temporelle (*timestamp*).

En d'autres termes, la latence est évaluée uniquement pour les tweets contribuant au gain, comme la différence entre la date de renvoi du tweet et le *timestamp* du premier tweet renvoyé du cluster auquel le tweet appartient.

Afin de comparer les systèmes, la latence est ensuite moyennée par profil, alors que les mesures orientées gain sont moyennée par jour et par profil.

Deux points importants sont à considérer, points malheureusement non mentionnés clairement dans le protocole d'évaluation et pour lesquels une analyse détaillée du code des outils d'évaluation a été nécessaire :

— pour les mesures *EG*, *nCG*, et *GMP*, quelle que soit la valeur $\Pi_i(t)$ pour un tweet, seul $\Theta(t)$ est considéré pour l'évaluation de G (équation 5.11). En d'autres termes, chaque tweet est renvoyé sur son jour d'émission pour évaluer les mesures *EG*, *nCG*, and *GMP*.

— les jours silencieux dépendent des clusters déjà retrouvés par les systèmes et sont donc dépendants des systèmes.

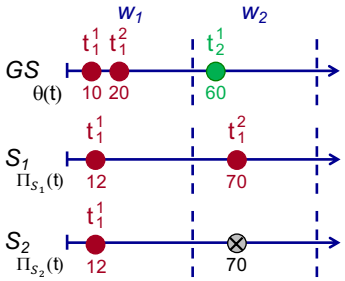


FIGURE 5.2 – Exemples de résultats renvoyés par deux systèmes S_1 et S_2 ainsi que la vérité terrain associée (GS), hypothèse H2.

Mesures	Systèmes	
	S_1	S_2
EG-0	$(\frac{1}{2} * 1 + 0)/2 = 0.25$	$(\frac{1}{1} * 1 + 0)/2 = 0.5$
EG-1	$(\frac{1}{2} * 1 + 0)/2 = 0.25$	$(\frac{1}{1} * 1 + 0)/2 = 0.5$
EG-p	$(\frac{1}{2} * 1 + 0)/2 = 0.25$	$(\frac{1}{1} * 1 + 0)/2 = 0.5$
Latence	2	2

TABLEAU 5.1 – Comportement des mesures EG et de latence par rapport à H1.

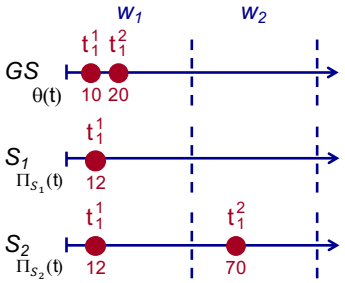


FIGURE 5.3 – Exemples de résultats renvoyés par deux systèmes S_1 et S_2 ainsi que la vérité terrain associée (GS), hypothèse H2.

Metrics	Systèmes	
	S_1	S_2
EG-0	$(\frac{1}{1} * 1 + 0)/2 = 0.5$	$(\frac{1}{2} * 1 + 0)/2 = 0.25$
EG-1	$(\frac{1}{1} * 1 + 1)/2 = 1$	$(\frac{1}{2} * 1 + 1)/2 = 0.75$
EG-p	$(\frac{1}{1} * 1 + 1)/2 = 1$	$(\frac{1}{2} * 1 + 1)/2 = 0.75$
Latency	2	2

TABLEAU 5.2 – Comportement des mesures EG et de latence par rapport à H2.

5.2.4.3 Analyse des hypothèses de la tâche

Deux hypothèses sont prises par les organisateurs :

— **H1.** un tweet redondant est équivalent à un tweet non pertinent

— **H2.** un score parfait sur les mesures EG-1 et nCG-1 peut être obtenu si le silence est respecté les jours silencieux.

Les contre-exemples présentés sur les figures 5.2 et 5.3 et issus de (Hubert et al., 2017a,b) réfutent ces hypothèses.

H1. En considérant les exemples de la figure 5.2 et du tableau 5.1, le système S_2 obtient de meilleurs scores que S_1 sur la mesure EG, alors que les deux résultats sont censés êtres équivalents (le premier tweet du cluster C_1 durant le jour w_1 et respectivement un tweet redondant et un tweet non pertinent durant le jour w_2). S_1 est plus pénalisé de renvoyer un tweet redondant qui sera ramené sur son jour d'émission qu'un tweet non pertinent, **H1** est donc invalidée.
H2. En considérant les exemples de la figure 5.3 et du tableau 5.2, w_2 est un jour silencieux pour les deux systèmes. S_2 brise le silence avec t_1^2 et obtient cependant un score parfait sur ce jour, comme S_1 qui ne renvoie aucun tweet, **H2** est donc invalidée. Ceci est une nouvelle fois dû au fait que les tweets sont ramenés sur leur fenêtre d'émission, quel que soit le jour auquel ils ont été réellement renvoyés.

5.2.4.4 Discussion complémentaire sur les paramètres de la tâche

Un des paramètres importants de la tâche est la limitation pour les systèmes à ne renvoyer au maximum que 10 tweets par jour. Afin de voir l'impact de cette limitation sur la mesure officielle $EG-1$, nous avons appliqué trois stratégies différentes sur les *runs* officiels de 2016 :

— dans la stratégie *First*, nous conservons les N premiers tweets selon la date de publication renvoyés chaque jour par les systèmes,

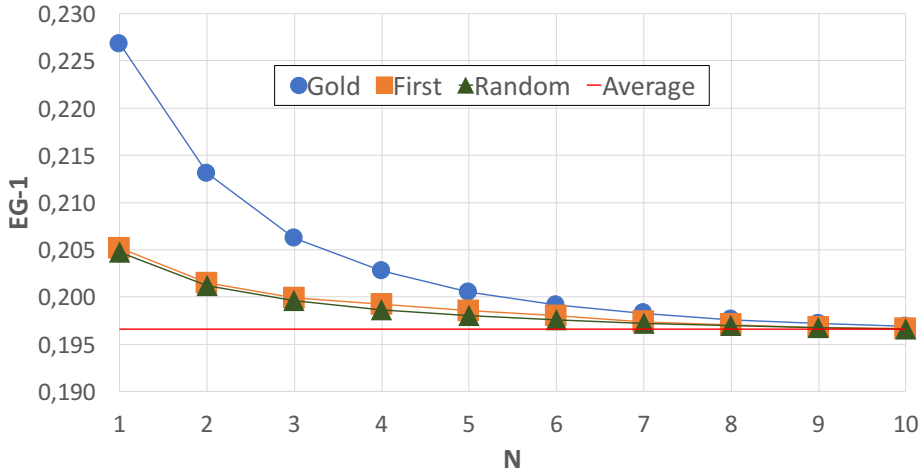


FIGURE 5.4 – Impact de la limite du nombre de tweets par jour sur la mesure *EG*-1, extrait de (Hubert et al., 2017b).

— dans la stratégie *Gold*, les N tweets conservés sont choisis pour maximiser $EG-1$,

— dans la stratégie *Random*, N tweets sont aléatoirement choisis.

La figure 5.4 montre les résultats de ces trois stratégies en faisant varier N de 1 à 10.

On voit clairement que pour obtenir de bons résultats sur $EG-1$, il est toujours préférable de renvoyer peu de tweets, ce qui renforce le fait qu'$EG-1$ est une mesure essentiellement orientée précision. De façon plus marquée encore, renvoyer un tweet par jour quelle que soit la stratégie choisie est systématiquement récompensé sur la mesure. La converture n'étant jamais évaluée dans la tâche, réussir à optimiser la mesure n'est pas si compliqué.

5.2.4.5 Conclusion

Les travaux présentés dans cette section sont détaillés dans (Hubert et al., 2017a,b). Voici ce qu'il est important d'en retenir :

— nous avons clarifié certains points liés au protocole d'évaluation de la tâche, points non définis clairement par les organisateurs. Seule une analyse approfondie du code d'évaluation nous a permis de les comprendre, nous laissant penser que peu de participants en sont conscients :

  — les tweets sont systématiquement ramenés sur leur jour d'émission pour l'évaluation (le jour auquel ils sont renvoyés par les systèmes importe peu).

  — les jours silencieux sont dépendants des systèmes, et il n'y a donc aucun sens à créer des approches cherchant à les détecter sans considérer les tweets déjà retournés.

— la couverture n'est pas évaluée par les mesures officielles. Il vaut toujours mieux pour un système renvoyer peu de tweets qui sont très probablement pertinents que d'essayer de trouver tous les tweets pertinents. Chercher à optimiser la couverture va probablement mener à une dégradation des résultats. Ce comportement des mesures avait été observé par les organisateurs de la tâche (Qian et al., 2016) mais ils avaient attribué cela à une mauvaise configuration des systèmes. Par conséquent, chaque article de recherche reportant des résultats de la tâche devrait systématiquement se comparer avec une *baseline* très simple renvoyant au plus un tweet par jour. Nos résultats sur la tâche 2017 confirment ces conclusions. Nous avons soumis un *run* très simple pour lequel nous renvoyons (au maximum) le premier tweet par jour contenant tous les termes du profil. Ce *run* nous a permis de nous classer 2ème sur l'évaluation temps-réel et 4ème sur l'évaluation *batch*, avec une latence parfaite (Lin et al., 2017).

— un dernier point (non détaillé ici) concerne la réutilisabilité de la collection avec les outils fournis par les organisateurs : un des fichiers fourni est incomplet, ce qui sur-évalue les systèmes. Il est malheureusement fort probable que de nombreux articles déjà publiés et utilisant ces collections reportent des résultats sur-évalués.

5.3 Synthèse des travaux présentés

Contributions principales. De façon synthétique, les contributions principales des travaux présentés dans ce chapitre sont les suivantes :

— des collections de test, que nous avons construites à partir de zéro afin de répondre à des problématiques de recherche spécifiques :
  — deux collections de tests en français pour l'évaluation du *clustering* de documents (un première de 2.6 millions de documents avec 25 requêtes mono-aspect, et une seconde de 10 millions de documents avec 25 requêtes mono-aspect et 25 requêtes multi-aspects). Il s'agit à notre connaissance de la seule collection de RI permettant d'évaluer des techniques de *clustering* de résultats ne respectant pas la *Cluster Hypothesis*.
  — deux (très) grandes collections de test pour la détection de copies d'images. La deuxième est à ce jour et à notre connaissance une des plus grandes collections utilisée pour l'évaluation de la recherche d'images.
— des protocoles d'évaluation
  — un protocole d'évaluation dans le cadre de tâches de recherche basées sur *clustering* de résultats ne respectant pas la *Cluster Hypothesis*,
  — un protocole d'évaluation pour la détection de copies d'images, notamment pour savoir si une image est une copie d'une autre,
— des études de protocoles d'évaluation sur des tâches de recherche liées à la recherche d'information agrégée :
  — nous avons montré que pour évaluer correctement les approches de recherche agrégée inter-verticale, il était préférable d'utiliser des requêtes avec besoins fixes et une évaluation de la pertinence par le contenu, ce qui n'avait jamais été clairement démontré, à notre connaissance, dans la littérature.
  — nous avons découvert des biais dans les protocoles d'évaluation TREC *Real-Time Summarization* 2016 et 2017, biais remettant également en question la réutilisabilité des collections et la validité des approches d'ores et déjà proposées dans la littérature pour ces tâches.

Diffusion scientifique.
La figure 5.5 résume en termes de publications les contributions détaillées dans ce chapitre, et récapitule toutes mes participations à des campagnes d'évaluation internationales.

Projets. Une partie de ces travaux a été menée dans le cadre du projet européen Quaero, pour lequel j'ai été responsable de deux workpackage Evaluation (*Document Ranking Optimization* et *Indexing Multimedia Objects*). Ce projet, mené de 2008 à 2013, m'a permis de collaborer avec la société Exalead, l'équipe CLLE-ERSS (Cognition, Langues, Langage, Ergonomie) de l'Université Toulouse 2, et l'équipe INRIA Texmex de Rennes.

Pour clôturer la liste de mes activités liées à l'évaluation des systèmes, j'ai été co-présidente, avec Jaap Kamps, Université d'Amsterdam, du comité de programme de la conférence CLEF 2015 [3] (Cappellato et al., 2015; Mothe et al., 2015). CLEF regroupe une conférence orientée vers l'évaluation et les campagnes d'évaluation, ainsi qu'un certain nombre de *labs*, chacun dédié à une tâche d'évaluation particulière

3. http://clef2015.clef-initiative.eu, dernier accès en mars 2018.

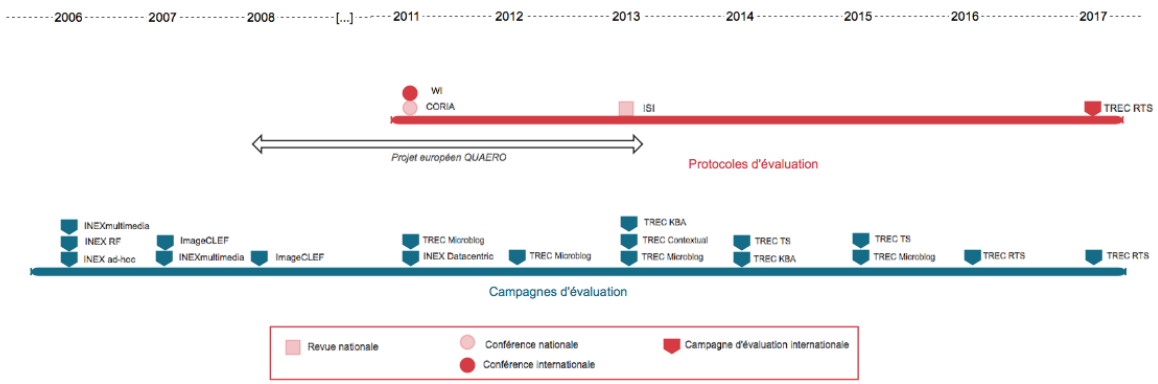


FIGURE 5.5 – Graphique synthétique de la structuration et valorisation des travaux sur le thème de l'Évaluation des systèmes

Conclusion

> [...] and the moral of that is — "Be what you would seem to be" — or if you'd like it put more simply — "Never imagine yourself not to be otherwise than what it might appear to others that what you were or might have been was not otherwise than what you had been would have appeared to them to be otherwise."
>
> — The Duchess

Ce mémoire a synthétisé l'essentiel des recherches que j'ai menées et encadrées entre 2006 et 2017, c'est-à-dire pour la période postérieure à mon doctorat. Mes travaux, situés dans le domaine de la Recherche d'Information, ont été présentés selon le cadre général de la recherche d'information agrégée. Deux parties du *framework* de la recherche d'information agrégée ont plus particulièrement été traitées : la phase de recherche de granules documentaires et la phase d'agrégation de l'information. Plus précisément, mes travaux ont concerné les axes suivants :

- la recherche de granules XML à l'aide de l'information structurelle,
- la recherche de granules de type microblogs et la gestion de l'information temps-réel,
- la recherche de granules autour des entités nommées, et plus particulièrement des relations autour des entités ou encore des documents vitaux liés aux entités dans le cas d'entités variant dans le temps,
- l'agrégation d'information autour des entités, sous forme de tableau relationnel ou encore de résumé temporel.

Un dernier axe, transversal aux précédents, concerne l'évaluation des approches de recherche d'information.

Cette conclusion présente tout d'abord un tableau résumant mes activités d'encadrement et mes publications, puis revient de façon synthétique sur les différents axes abordés.

Graphique synthétique de la valorisation des travaux

Le graphique de la Figure 5.6 résume de façon chronologique mes publications et encadrements de thèses/masters. La majorité de cette production scientifique est le résultat de collaborations et travaux menés en équipe. J'ai participé au co-encadrement de 8 étudiants en thèse ; les contributions issues de ces recherches ont été très largement détaillées dans ce mémoire.

Le détail de ces publications (avec les indicateurs bibliométriques associés), encadrements et participations à des projets se trouve dans mon CV en Annexe B de ce document. On y trouvera également mes activités et implications scientifiques de nature collective (participation à des comités de programme, organisations de conférence, implications dans la vie scientifique nationale).

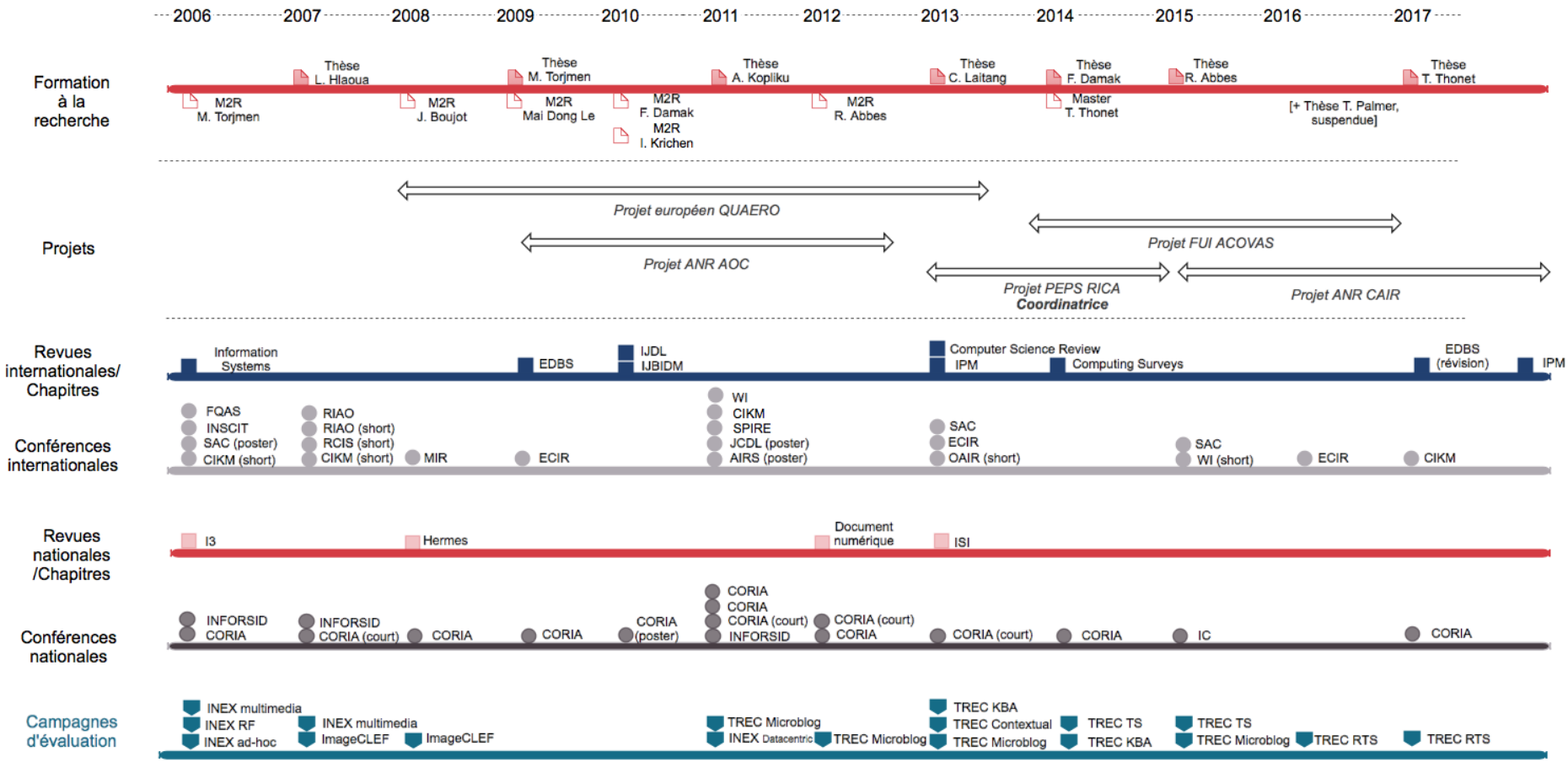


FIGURE 5.6 – Tableau synthétique de mes encadrements de thèses et masters, publications et participations à des projets.

Synthèse des travaux

Les travaux présentés dans ce mémoire concernent deux thématiques générales de la Recherche d'Information : l'appariement documents/requêtes (par l'adaptation de modèles existants à nos problématiques ou par la construction de nouveaux modèles), et l'évaluation (par la proposition de protocoles d'évaluation). Les contributions ont été à la fois théoriques et expérimentales. Nous les synthétisons ici par rapport aux différents axes de recherche que nous avons considérés.

Axe 1 : Recherche de granules XML. Les recherches menées dans cet axe sont dans la continuité directe de mes travaux de thèse en RI structurée. Nous avons étudié l'impact de la structure des documents sur le processus de recherche de granules XML à plusieurs niveaux : pour répondre à des requêtes pas ou peu structurées, pour la réécriture de requêtes ou encore pour la recherche de granules multimédia. La conclusion principale de ces travaux est que la structure est un indicateur très fort de pertinence : nous avons montré qu'il existe une relation contextuelle entre les termes des documents et les structures pertinentes, et que la structure doit être prise en compte de façon explicite pour répondre aux requêtes très structurées.

J'ai co-encadré 3 thèses sur ces thématiques (Hlaoua, 2007; Laitang, 2013; Torjmen, 2009), et une partie de ces recherches a été menée dans le projet ANR AOC (*Appariement d'Objets Complexes*, 2009-2012).

Axe 2 : Recherche de granules de type microblog. Mes travaux dans ce cadre ont principalement porté sur l'identification des caractéristiques pouvant traduire la pertinence des microblogs. Nous avons montré l'importance des facteurs liés aux URLs présentes dans les tweets, mais aussi le manque d'intérêt de facteurs pourtant couramment utilisés dans l'état de l'art, comme les facteurs liés aux hashtags et à la notion de temps-réel. Je continue actuellement à travailler sur cette thématique, en approfondissant la dimension temps-réel de la recherche et surtout les protocoles d'évaluation nécessaires dans ce cadre.

J'ai co-encadré 2 thèses sur ces thématiques [4] (Damak, 2014), et une partie de ces travaux a été réalisée dans le cadre du projet FUI ACOVAS (*Outils Agile pour la COnception et VAlidation Système*, 2014-2016).

Axe 3 : Recherche de granules autour des entités. Nous avons travaillé sur deux problématiques différentes de la recherche de granules autour des entités : la recherche de relations entités-entités et entités-attributs, ainsi que la recherche de documents vitaux (frais) pour les entités. La source d'information considérée a été dans ces deux cas le Web : dans notre premier cas d'étude dans un scénario de recherche *adhoc* en extrayant les tables relationnelles ou les listes des pages HTML, et dans notre second cas d'étude dans un scénario de filtrage d'information pour trouver de l'information nouvelle sur les entités.

Axe 4 : Agrégation d'information autour des entités. Une fois les granules d'information extraits pour une requête (documents, granules XML, microblogs, relations, ou encore images et vidéos), la phase finale de la recherche d'information agrégée consiste à construire l'agrégat. Toujours dans un scénario de recherche autour d'une entité nommée ou d'une classe d'entités, nous avons proposé d'agréger les informations à travers soit (i) d'un tableau relationnel, soit (ii) d'un résumé temporel autour de l'entité. Ce résumé temporel pourra également servir à la mise à jour de bases de connaissances, dont le délai de mise à jour peut parfois atteindre une année.

4. Dont une n'a pas été menée à son terme.

Nous avons, à travers les axes 3 et 4, donné deux scénarios complets de recherche d'information agrégée.

Le premier, résumé sur la figure 5.7 concerne la **recherche agrégée relationnelle**. Dans ce cadre, la requête peut être une requête de type attribut, entité ou classe d'entités. Sur notre exemple, la requête `Filmographie des oeuvres de Lewis Carroll` est une requête classe pour laquelle nous commençons à identifier les entités instances de la classe à partir d'une base de connaissance telle que DBPedia ou Wikipedia. Notre méthode de recherche des attributs représentatifs de ces entités est ensuite appliquée (section 3.2.1). Ces attributs sont finalement triés par pertinence à la requête et présentés sous forme de tableau relationnel à l'utilisateur (section 4.2.1). Dans notre exemple, les attributs `titre`, `année`, `producteur`, `réalisateur`, `description`, `affiche du film` et `opinions` ont été retenus.

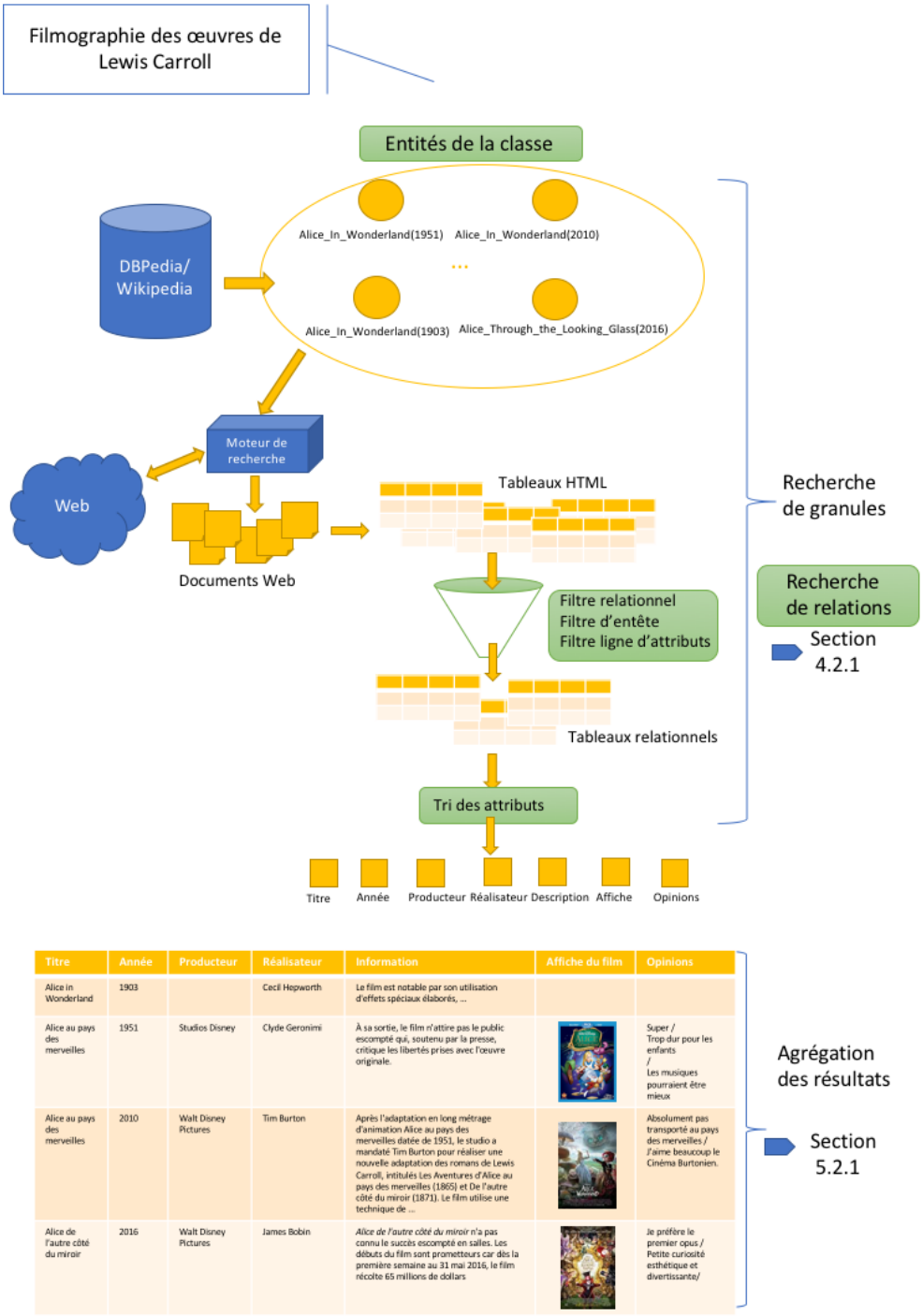
Notre second scénario, résumé sur la Figure 5.8 concerne le **résumé temporel d'information** autour d'une entité. Dans ce scénario, nous partons d'une entité nommée (ici `Anne Hathaway`), et d'une connaissance a priori sur l'entité (par exemple son profil Wikipedia) à un instant t (ici `30/1/2010`). Le but est de proposer à l'utilisateur un résumé temporel contenant des informations nouvelles sur l'entité, venant compléter le profil déjà connu. Une première étape (section 3.2.2) filtre à partir du flux du Web les documents apportant des informations nouvelles (documents vitaux). Nous avons proposé deux méthodes différentes, basées sur des modèle de langue de vitalité ou bien sur l'utilisation d'expressions temporelles. L'étape suivante (section 4.2.2) consiste à extraire de ces documents vitaux les phrases qui construiront le résumé temporel (dans notre exemple ces phrases donnent des informations sur un rôle décroché par `Anne Hathaway` dans un film de `Tim Burton` au début de l'année `2010`). Pour ce faire, nous avons proposé de sélectionner les phrases à partir de certains mots déclencheurs et d'une méthode de détection de la nouveauté utilisant les entités liées.

Ces deux axes ont donné lieu au co-encadrement de deux thèses (Abbes, 2015; Kopliku, 2011). Une partie des travaux relatifs à ces axes a également été menée dans le cadre du projet PEPS RICA (*Recherche d'information en Contexte et Agrégée*, 2013-2014) et du projet ANR CAIR (*Contextual Aggregated Information Retrieval*, 2015-2018).

Axe 5 : Évaluation. Tous les travaux présentés dans ce mémoire ont été systématiquement évalués sur des collections de test du domaine (INEX, CLEF, TREC). Pour les rares cas où ces collections n'existaient pas, nous avons construit nos propres *benchmarks* d'évaluation (recherche agrégée relationnelle). J'ai également participé dans le cadre du projet européen Quaero (2008-2012) à la construction de deux collections de test complètes (documents, requêtes, jugements de pertinence, protocoles d'évaluation). Certaines de mes recherches ont enfin concerné les protocoles d'évaluation, notamment celui du résumé temps-réel de tweets.

Travaux connexes. Une petite partie des recherches que j'ai menées ou encadrées n'a pas été mentionnée dans le document, et je profite de cette conclusion pour les synthétiser rapidement.

— Les travaux de Thibaut Thonet (Thonet et al., 2016, 2017), dont j'ai co-encadré la thèse avec Guillaume Cabanac et Mohand Boughanem (Thonet, 2017), ont porté sur l'utilisation de modèles thématiques pour la découverte de points de vue sur le Web. De nombreux travaux de la littérature s'intéressent à l'analyse de sentiments et la fouille d'opinions. De telles recherches pourraient être utilisées dans notre cas pour synthétiser les opinions (un cas d'application est par exemple la colonne `Opinions` du tableau agrégat de la Figure 5.7). Ces travaux se focalisent cependant sur des opinions simplement positives ou négatives. Les travaux de Thibaut se sont quant à eux intéressés aux points de vue, qui généralisent l'opinion au delà de son acception usuelle liée à la polarité (positive ou négative) et permettent l'étude d'opinions exprimées plus subtilement, telles que les opinions politiques. L'idée a été d'utiliser des modèles thé-



Titre	Année	Producteur	Réalisateur	Information	Affiche du film	Opinions
Alice in Wonderland	1903		Cecil Hepworth	Le film est notable par son utilisation d'effets spéciaux élaborés, ...		
Alice au pays des merveilles	1951	Studios Disney	Clyde Geronimi	À sa sortie, le film n'attire pas le public escompté qui, soutenu par la presse, critique les libertés prises avec l'œuvre originale.		Super / Trop dur pour les enfants / Les musiques pourraient être mieux
Alice au pays des merveilles	2010	Walt Disney Pictures	Tim Burton	Après l'adaptation en long métrage d'animation Alice au pays des merveilles datée de 1951, le studio a mandaté Tim Burton pour réaliser une nouvelle adaptation des romans de Lewis Carroll, intitulés Les Aventures d'Alice au pays des merveilles (1865) et De l'autre côté du miroir (1871). Le film utilise une technique de ...		Absolument pas transporté au pays des merveilles / J'aime beaucoup le Cinéma Burtonien.
Alice de l'autre côté du miroir	2016	Walt Disney Pictures	James Bobin	*Alice de l'autre côté du miroir* n'a pas connu le succès escompté en salles. Les débuts du film sont prometteurs car dès la première semaine au 31 mai 2016, le film récolte 65 millions de dollars		Je préfère le premier opus / Petite curiosité esthétique et divertissante/

FIGURE 5.7 – Scénario basé sur la recherche d'information agrégée relationnelle

matiques (*Topic models*) pour découvrir simultanément les thèmes et les points de vue exprimés dans des corpus de textes d'opinion. Une première contribution a été de différencier mots d'opinions (spécifiques à la fois à un point de vue et à un thème) et mots thématiques (dépendants du thème mais neutres vis-à-vis des différents points de vue) en nous basant sur les parties de discours (Thonet et al., 2016). Une seconde contribution a concerné les réseaux sociaux, où l'objectif était d'analyser dans quelle mesure l'utilisation des interactions entre utilisateurs, en outre de leur contenu textuel généré, est bénéfique à l'identification de leurs points de vue (Thonet et al., 2017).

— J'ai également participé à la définition d'une approche pour la recherche d'information contextuelle (Hubert et al., 2013; Palacio et al., 2013) dans le cadre de travaux menés

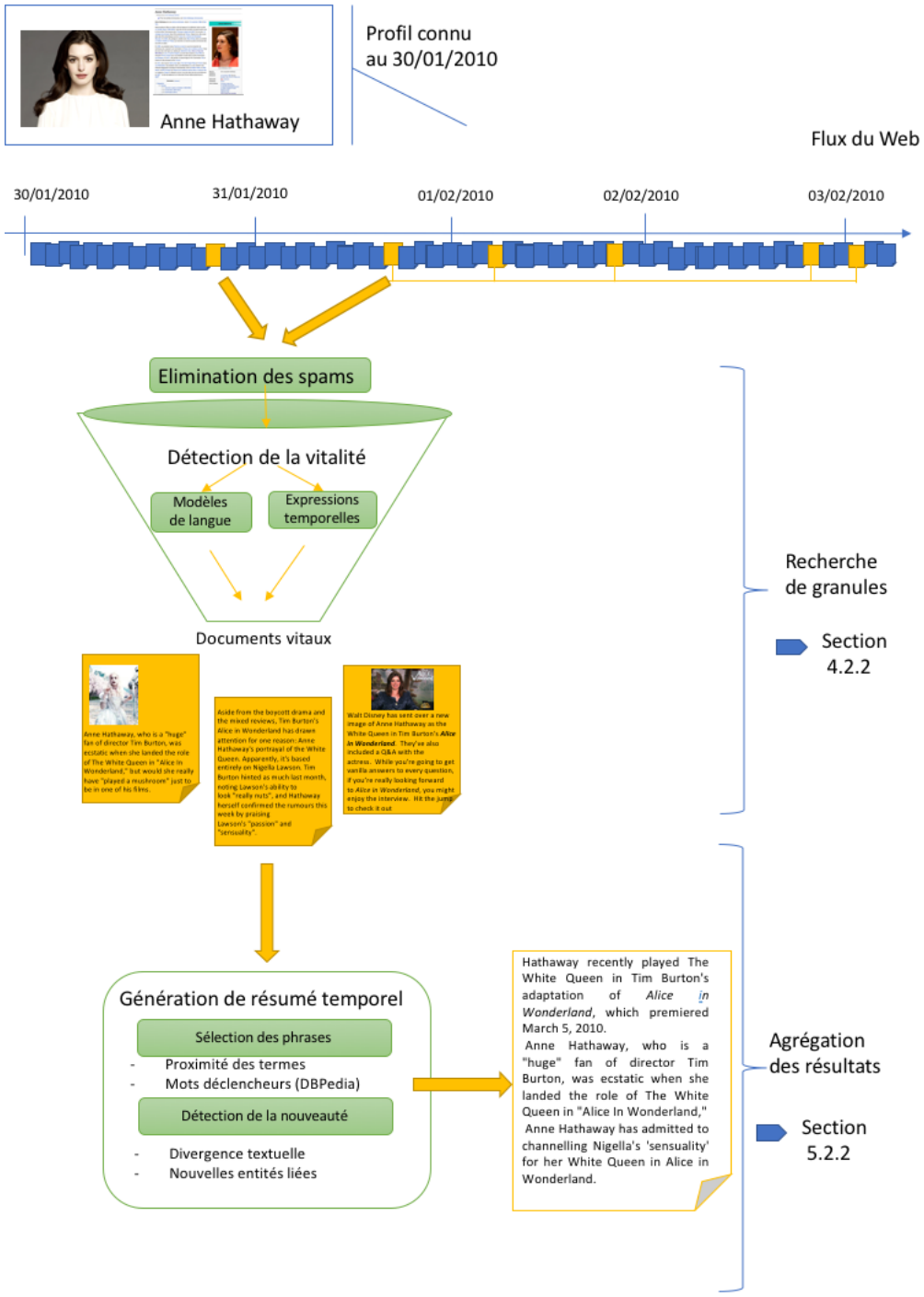


FIGURE 5.8 – Scénario basé sur la recherche d'information agrégée autour des entités variant dans le temps

avec le LIUPPA (Pau). Ces travaux cherchent à répondre à des besoins utilisateurs géolocalisés, le but est de renvoyer des lieux à visiter selon des profils utilisateurs donnés. L'approche que nous avons proposée combine des services et outils de RI géographique (Google Places [5], Yahoo ! BOSS Geo Services [6], PostGIS [7], Gisgraphy [8], GeoNames [9]) et

5. https://developers.google.com/places, dernier accès mars 2018
6. https://developer.yahoo.com/boss/geo/, dernier accès mars 2018
7. https://postgis.net, dernier accès mars 2018
8. http://www.gisgraphy.com, dernier accès mars 2018
9. http://www.geonames.org, dernier accès mars 2018

trie les résultats selon des caractéristiques telles que la distance de la proposition/lieu résultat à l'utilisateur, sa popularité et bien sûr les préférences de l'utilisateur. Nous avons participé avec cette approche à la campagne d'évaluation TREC *Contextual Suggestion Track* 2013.

— Enfin, je travaille actuellement, avec Gilles Hubert, Yoann Pitarch et Ronan Tournier, à la définition d'un nouveau modèle de recherche d'information, Tournarank (Hubert et al., 2018a,b), sur lequel je vais revenir dans les perspectives de ce mémoire.

Perspectives

> "Would you tell me, please, which way I ought to go from here?"
> "That depends a good deal on where you want to get to," said the Cat.
> "I don't much care where–" said Alice.
> "Then it doesn't matter which way you go," said the Cat.
> "–so long as I get *somewhere*," Alice added as an explanation.
> "Oh, you're sure to do that," said the Cat, "if you only walk long enough."

Mes perspectives de recherche se situent naturellement dans le prolongement des recherches présentées dans ce mémoire. J'envisage deux axes principaux pour mes recherches futures.

Axe 1 : Construction de l'agrégat. Nos premières recherches sur la construction de l'agrégat se sont focalisées sur des scénarios spécifiques liés aux entités (recherche agrégée relationnelle et résumé temporel). Dans un premier temps, je souhaiterais **proposer des solutions et approches pour d'autres scénarios de construction d'agrégats**. Nous avons récemment défini un nouveau modèle de RI, Tournarank (Hubert et al., 2018a,b), dont le but est d'utiliser les caractéristiques des documents pour évaluer leur pertinence finale. De nombreuses approches utilisant les caractéristiques des documents existent dans la littérature pour l'ordonnancement des résultats, mais elles requièrent presque toutes une phase d'apprentissage (on peut par exemple citer les méthodes de *Learning To Rank* (Liu, 2011)). L'approche que nous proposons est non supervisée, et elle propose d'ordonner les documents en s'inspirant des compétitions sportives. La figure 5.9 décrit notre proposition : les documents sont représentés par un ensemble de caractéristiques (partie (a) de la figure) et s'affrontent lors de tournois. Un tournoi est vu comme une séquence de matchs au cours desquels deux documents s'affrontent sur la base des valeurs de leurs caractéristiques. Ces caractéristiques peuvent être la fréquence des termes de la requête dans le document, son score selon un modèle de RI, sa longueur... ; elles sont variables d'une collection à l'autre. À l'issue du tournoi, les documents sont ordonnés dans l'ordre décroissant des scores obtenus durant le tournoi (partie (b)). Durant chaque match, les documents s'affrontent en comparant deux à deux les valeurs de leurs caractéristiques. Chaque document possède une jauge de vie qui sera décrémentée à chaque "coup" reçu du document adverse. Le match s'arrête lorsque tous les documents ont joué (partie (c)). Le principal point fort de notre approche est qu'elle est entièrement paramétrable (type de tournoi, règle des matchs, etc) et généralisable.

Mon objectif est donc de m'en servir comme base pour utiliser en entrée des documents/granules d'informations hétérogènes (images, videos, news, microblogs, texte,...), dé-

crits par des caractéristiques qui leur sont propres. Au lieu de faire s'affronter les documents deux à deux, les documents pourraient être regroupés en équipes selon des critères de complétude, non redondance, diversité, etc. Les équipes (qui sont donc ici assimilées à des agrégats) pourront ensuite s'affronter les unes les autres pour déterminer l'agrégat final qui sera présenté à l'utilisateur. Un scénario possible d'utilisation de l'approche est celui du résumé d'information sur Twitter : devant les millions de tweets parfois publiés sur un événement, il est difficile pour l'utilisateur de s'y retrouver avec une simple recherche par *hashtags*. L'idée serait de présenter un résumé de l'événement, composé à la fois d'images, de news, de vidéos et de tweets.

De façon plus générale, et comme nous l'avons déjà soulevé, synthétiser l'information soulève différentes problématiques liées à la recherche d'information, notamment relatives à :

— la redondance d'information,

— la multiplicité des types d'information (texte, images, vidéos...),

— la complémentarité des éléments d'information utilisés en termes de thématiques abordées, d'éléments spatiaux et temporels apportés, par exemple,

— la validité, la crédibilité des éléments d'information manipulés,

— la construction et la restitution de l'information synthétisée sous forme structurée,

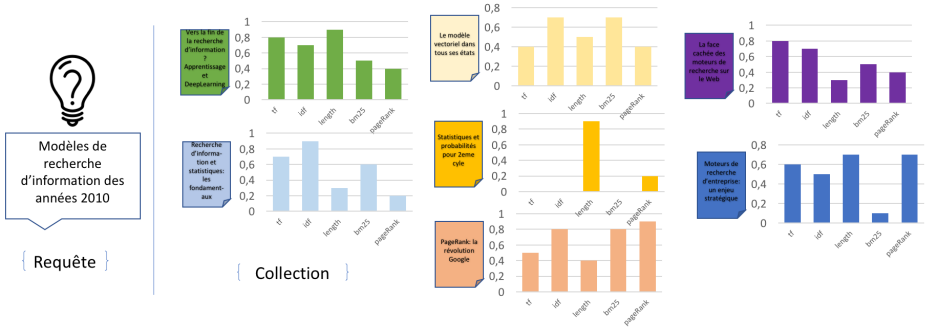
— l'éventuelle dimension temps-réel de l'information.

Nos travaux traitent certaines de ces problématiques, mais aucun de nos modèles ne les considère toutes et dans leur ensemble. À plus long terme et indépendamment de tout scénario, je souhaiterais donc **généraliser les approches existantes et traduire ces problématiques de construction de l'agrégat grâce à des critères objectifs que doit respecter l'agrégat**. Cela passera dans un premier temps par la définition des critères, puis par leur combinaison pour définir les agrégats. Chaque critère peut traduire une propriété de granule, des relations entre granules, etc. On peut citer par exemple la pertinence d'un granule vis-à-vis de la requête, sa nouveauté, la diversité des granules entre eux, la redondance, complémentarité et nouveauté d'un granule vis-à-vis des autres, la cohésion entre les granules, leur compatibilité ou encore la taille du résultat final. Une fonction objective prenant en compte ces critères pourrait ensuite être utilisée pour construire les agrégats, indépendamment du scénario de recherche d'information considéré.

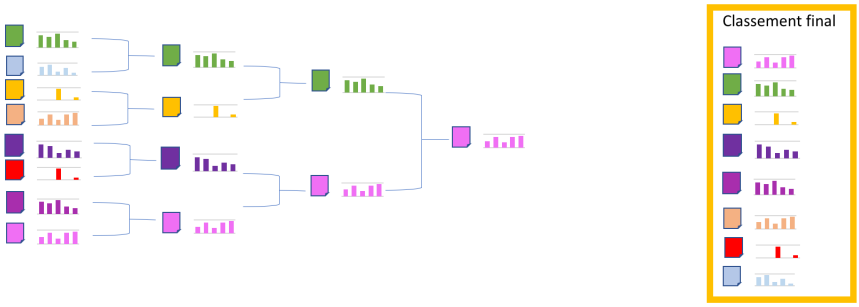
Axe 2 : Évaluation des approches. Confortée par mon expérience acquise sur le montage de collections de tests et de protocoles d'évaluation, je souhaite poursuivre mes recherches dans le domaine en travaillant selon deux directions :

1. je souhaite tout d'abord travailler sur la **validité des protocoles de tests actuels**. Les tâches de RI se diversifient, et les protocoles d'évaluation du type Cranfield (Cleverdon, 1967; Sanderson, 2010a) pourtant encore très utilisés dans des campagnes telles que TREC ou CLEF commencent aujourd'hui à montrer leurs limites (Fuhr, 2017). Les collections de documents ne sont aujourd'hui plus statiques. Le même constat peut être fait sur les besoins utilisateurs. Les approches de RI actuelles intègrent la dimension-temps réel de l'information et du besoin (flux du Web, flux Twitter) et les protocoles de tests doivent s'adapter. On voit ainsi apparaître de nombreuses nouvelles tâches de recherche aux sein des campagnes d'évaluation, comme la tâche *TREC Open Search* (Balog et al., 2016) pour laquelle les jugements de pertinence sont faits en temps-réel par de vrais utilisateurs ou encore la tâche *TREC Dynamic Domain Track* (Yang and Soboroff, 2016) dont le but est de produire des algorithmes de recherche interactifs s'adaptant à un besoin en information dynamique.

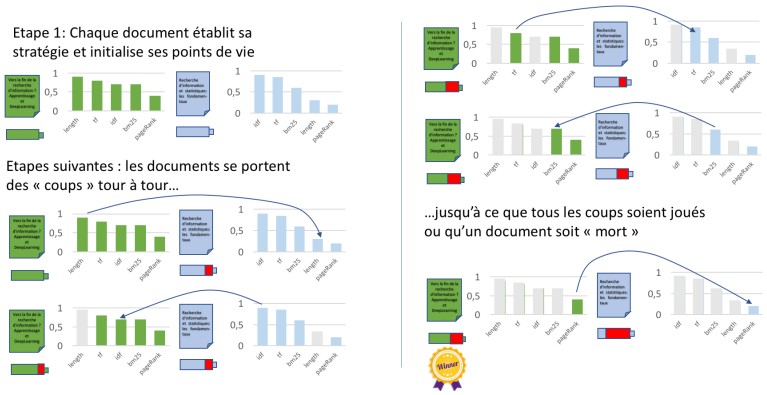
   Nos recherches récentes sur la tâche *Real-Time Summarization* (RTS) ont montré un biais important dans le protocole d'évaluation (Hubert et al., 2017a,b). La communauté de RI est bien consciente de ce manquement des nouvelles tâches d'évaluation proposées,



(a) Représentation des documents sous forme d'ensemble de caractéristiques (*features*)



(b) Les documents sont ordonnés selon les résultats de matchs les opposants, matchs organisés dans le cadre de tournois



(c) Exemple de fonctionnement d'un match

FIGURE 5.9 – Fonctionnement de notre approche Tournarank basée sur des matchs entre documents organisés sous forme de tournois. Les documents sont représentés à l'aide de caractéristiques.

en témoignent les recherches menées dans le cadre de la tâche *TREC Common Core Track* (Allan et al., 2017) : l'idée est d'établir de nouvelles collections de test fiables et surmontant les limites de l'évaluation de type Cranfield. Je souhaiterais me baser sur les travaux récents de la RI axiomatique (Amigo et al., 2017; Busin and Mizzaro, 2013) pour passer au crible ces nouveaux protocoles d'évaluation. Cela permettra soit de démontrer leur validité, soit de trouver de nouveaux biais et proposer par là même de nouveaux protocoles basés sur les axiomes que les protocoles d'évaluation devront respecter.

2. une autre piste de recherche concerne naturellement l'**évaluation de la recherche d'information agrégée**. Un certain nombre de tâches dans les campagnes d'évaluations internationales peuvent y être reliées : la tâche de tri d'entités d'INEX (*Entity Ranking Track*) (Demartini et al., 2009), la tâche Entité de TREC (*Entity Track*) (Balog et al., 2010), le challenge SemSearch évaluant la recherche d'entités dans les *Linked Data* (Tran et al., 2011), ou encore les tâches TAC-KBP (*Knowledge Base Population*) (Ji et al., 2014) et TREC KBA (*Knowledge Base Acceleration*) (Frank et al., 2013). Ces initiatives sont principalement relatives à la recherche d'information textuelle et aux entités, mais ne traitent pas directement du problème d'agrégation. La tâche *Temporal Summarization* de TREC (avec le biais que nous avons mentionné au chapitre 5), est l'une des seules à évaluer la qualité de l'objet agrégé dans sa globalité. Une autre initiative notable concerne la tâche CAR (*Complex Answer Retrieval Track*) proposée à TREC en 2017. L'idée est de produire des résultats synthétiques (semblables à des pages Wikipedia) agrégeant des granules provenant de documents différents du corpus. Le protocole d'évaluation de cette tâche est à ce jour le plus proche de ce qu'il faudrait pour évaluer la recherche d'information agrégée dans sa globalité. La dimension temps-réel de l'information y est pourtant passée sous silence.

   Les protocoles de tests proposés pour la recherche agrégée devront se soucier de la qualité et de la nouveauté du résultat final, de la cohérence et de la diversité des granules qui le composent ainsi que de la qualité des relations potentielles entre eux. Ceci ne pourra pas se faire sans de nouvelles mesures d'évaluation (éventuellement proposées elles aussi via des axiomes, cf. point 1) ainsi que sans le recours à des évaluations utilisateurs intensives, via du *crowdsourcing* (Alonso et al., 2008; Zhao and Zhu, 2014).

Bibliographie

Rafik Abbes. Étude de l'apport du Web de données et du Web relationnel dans la recherche agrégée. Rapport de master, IRIT, Université Paul Sabatier, Toulouse, Juin 2012. *Cité page 52*

Rafik Abbes. *Filtrage et agrégation d'informations vitales relatives à des entités*. Thèse de doctorat, Université Paul Sabatier, Toulouse, France, décembre 2015. https://tel.archives-ouvertes.fr/tel-01266560/, Soutenance le 11/12/2015. *5 citations pages 42, 43, 52, 69, et 90*

Rafik Abbes, Arlind Kopliku, Karen Pinel-Sauvagnat, Nathalie Hernandez, and Mohand Boughanem. Apport du Web et du Web de Données pour la recherche d'attributs. In *Conférence francophone en Recherche d'Information et Applications (CORIA), Neuchâtel, Suisse*, pages 37–46. Université de Neuchâtel, avril 2013a. http://doi.org/10.24348/coria.2013.coria2013_91. *3 citations pages iv, 46, et 47*

Rafik Abbes, Karen Pinel-Sauvagnat, Nathalie Hernandez, and Mohand Boughanem. IRIT at TREC Knowledge Base Acceleration 2013 : Cumulative Citation Recommendation Task. In *Text REtrieval Conference (TREC), Gaithersburg, USA*, http://www-nlpir.nist.gov/, novembre 2013b. National Institute of Standards and Technology (NIST). http://trec.nist.gov/pubs/trec22/papers/IRIT-kba.pdf. *4 citations pages 43, 52, 71, et V*

Rafik Abbes, Karen Pinel-Sauvagnat, Nathalie Hernandez, and Mohand Boughanem. IRIT at TREC Temporal Summarization 2014. In *Text REtrieval Conference (TREC), Gaithersburg, USA*, http://www-nlpir.nist.gov/, novembre 2014a. National Institute of Standards and Technology (NIST). http://trec.nist.gov/pubs/trec23/papers/pro-IRIT_ts.pdf. *3 citations pages 68, 71, et V*

Rafik Abbes, Karen Pinel-Sauvagnat, Nathalie Hernandez, and Mohand Boughanem. IRIT at TREC KBA 2014. In *Text REtrieval Conference (TREC), Gaithersburg, USA*, http://www-nlpir.nist.gov/, novembre 2014b. National Institute of Standards and Technology (NIST). http://trec.nist.gov/pubs/trec23/papers/pro-IRIT_kba.pdf. *4 citations pages 51, 52, 71, et V*

Rafik Abbes, Karen Pinel-Sauvagnat, Nathalie Hernandez, and 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), Nancy*, pages 129–143. LORIA, mars 2014c. http://doi.org/10.24348/sdnri.2014.CORIA-12. *3 citations pages 48, 51, et V*

Rafik Abbes, Nathalie Hernandez, Karen Pinel-Sauvagnat, and Mohand Boughanem. Détection d'informations vitales pour la mise à jour de bases de connaissances . In *Journées Francophones d'Ingénierie des Connaissances (IC), Rennes*, pages 147–158. Association Française d'Intelligence Artificielle (AFIA), juillet 2015a. https://hal.inria.fr/hal-01165507v1. *4 citations pages v, 64, 67, et V*

Rafik Abbes, Nathalie Hernandez, Karen Pinel-Sauvagnat, and Mohand Boughanem. 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*, pages 49–58. IEEE, décembre 2015b. http://doi.org/10.1109/WI-IAT.2015.32. *2 citations pages 64 et V*

Rafik Abbes, Bilel Moulahi, Abdelhamid Chellal, Karen Pinel-Sauvagnat, Nathalie Hernandez, Mohand Boughanem, Lynda Tamine, and Sadok Ben Yahia. IRIT at TREC Temporal Summarization 2015. In *Text REtrieval Conference (TREC), Gaithersburg, Maryland USA*, http://www.nist.org, novembre 2015c. National Institute of Standards and Technology (NIST). http://trec.nist.gov/act_part/conference/papers/IRIT-TS.pdf. *3 citations pages 68, 71, et V*

Rafik Abbes, Karen Pinel-Sauvagnat, Nathalie Hernandez, and Mohand Boughanem. Leveraging temporal expressions to filter vital documents related to an entity . In *ACM Symposium on Applied Computing (SAC), Salamanca, Spain*, pages 1093–1098. ACM, avril 2015d. http://dx.doi.org/10.1145/2695664.2695910. *3 citations pages 48, 51, et V*

Rafik Abbes, Karen Pinel-Sauvagnat, Nathalie Hernandez, and Mohand Boughanem. When temporal expressions help to detect vital documents related to an entity. *Applied Computing Review*, 15(3) :49–58, septembre 2015e. http://doi.org/10.1145/2835260.2835263. *3 citations pages 48, 51, et V*

Eugene Agichtein and Luis Gravano. Snowball : extracting relations from large plain-text collections. In *DL '00 : Proc. of the fifth ACM conference on Digital libraries*, pages 85–94, 2000. *Cité page 41*

Enrique Alfonseca, Marius Pasca, and Enrique Robledo-Arnuncio. Acquisition of instance attributes via labeled and related instances. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 58–65. ACM, 2010. *Cité page 40*

Abdelslam Alilaouar. *Contribution à l'interrogation flexible de données semi-structurées.* PhD thesis, Université Paul Sabatier, Toulouse, 2007. *Cité page 14*

James Allan, Donna Harman, Evangelos Kanoulas, Dan Li, Christophe Van Gysel, and Ellen Vorhees. Trec 2017 common core track overview. In *Text REtrieval Conference (TREC), Gaithersburg, Maryland, USA*, 2017. *Cité page 97*

Omar Alonso, Daniel E Rose, and Benjamin Stewart. Crowdsourcing for relevance evaluation. *ACM SigIR Forum*, 42(2) :9–15, 2008. https://doi.org/10.1145/1480506.1480508. *Cité page 97*

Sihem Amer-Yahia, Francesco Bonchi, Carlos Castillo, Esteban Feuerstein, Isabel Méndez-Díaz, and Paula Zabala. Complexity and algorithms for composite retrieval. In *Proceedings of the 22nd International Conference on World Wide Web*, pages 79–80. ACM, 2013. *Cité page 55*

Enrique Amigo, Hui Fang, Stefano Mizzaro, and ChengXiang Zhai. Axiomatic thinking for information retrieval : And related tasks. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1419–1420. ACM, 2017. *Cité page 97*

Gabor Angeli, Melvin Jose Johnson Premkumar, and Christopher D Manning. Leveraging linguistic structure for open domain information extraction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1 : Long Papers)*, volume 1, pages 344–354, 2015. *Cité page 42*

Jaime Arguello. Aggregated search. *Foundations and Trends in Information Retrieval*, 10(5) :365–502, 2017. http://dx.doi.org/10.1561/1500000052. *2 citations pages 2 et 4*

Jaime Arguello and Rob Capra. The effects of aggregated search coherence on search behavior. *ACM Transactions on Information Systems (TOIS)*, 35(1) :2, 2016. https://doi.org/10.1145/2935747. *Cité page 78*

Jaime Arguello, Fernando Diaz, Jamie Callan, and Jean-Francois Crespo. Sources of evidence for vertical selection. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 315–322. ACM, 2009. *2 citations pages 54 et 78*

Jaime Arguello, Fernando Diaz, and Milad Shokouhi. Integrating and ranking aggregated content on the web. In *WWW 2012, Tutorial, Lyon*, 2012. http://ils.unc.edu/~jarguell/www12_content_agg/. *Cité page 4*

Javed Aslam, Matthew Ekstrand-Abueg, Virgil Pavlu, Fernando Diaz, and Tetsuya Sakai. Trec 2013 temporal summarization. In *Text REtrieval Conference (TREC), Gaithersburg, Maryland, USA*, 2013a. *4 citations pages 42, 59, 67, et V*

Javed Aslam, Fernando Diaz, Matthew Ekstrand-Abueg, Richard McCreadie, Virgil Pavlu, and Tetsuya Sakai. Trec 2014 temporal summarization track overview. In *Text REtrieval Conference (TREC), Gaithersburg, Maryland, USA*, 2015. *5 citations pages 42, 59, 67, 82, et V*

Javed A Aslam, Matthew Ekstrand-Abueg, Virgil Pavlu, Fernando Diaz, and Tetsuya Sakai. Trec 2013 temporal summarization. In *Text REtrieval Conference (TREC), Gaithersburg, Maryland, USA*, 2013b. *Cité page 71*

Farzindar Atefeh and Wael Khreich. A survey of techniques for event detection in twitter. *Computational Intelligence*, 31(1) :132–164, 2015. https://doi.org/10.1111/coin.12017. *Cité page 29*

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. Dbpedia : A nucleus for a web of open data. In *The semantic web*, pages 722–735. Springer, 2007. *2 citations pages 45 et 47*

Yonatan Aumann, Ronen Feldman, Yair Liberzon, Benjamin Rosenfeld, and Jonathan Schler. Visual information extraction. *Knowl. Inf. Syst.*, 10 :1–15, July 2006. http://doi.org/10.1007/s10115-006-0014-x. *Cité page 42*

Krisztian Balog and Heri Ramampiaro. Cumulative citation recommendation : Classification vs. ranking. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pages 941–944. ACM, 2013. *Cité page 43*

Krisztian Balog, Arjen P. de Vries, Pavel Serdyukov, Paul Thomas, and Thijs Westerveld. Overview of the TREC 2009 Entity Track. In *Text REtrieval Conference (TREC), Working Notes, Gaithersburg, Maryland, USA*. NIST, November 2009. *2 citations pages 55 et 62*

Krisztian Balog, Pavel Serdyukov, and Arjen P de Vries. Overview of the trec 2010 entity track. In *Text REtrieval Conference (TREC), Gaithersburg, Maryland, USA*, 2010. *Cité page 97*

Krisztian Balog, Heri Ramampiaro, Naimdjon Takhirov, and Kjetil Nørvåg. Multi-step classification approaches to cumulative citation recommendation. In *Proceedings of the 10th Conference on Open Research Areas in Information Retrieval*, pages 121–128, 2013. *Cité page 43*

Krisztian Balog, Liadh Kelly, and Anne Schuth. Head first : Living labs for ad-hoc search evaluation. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pages 1815–1818. ACM, 2014. *Cité page 71*

Krisztian Balog, Anne Schuth, Peter Dekker, Narges Tavakolpoursaleh, Philipp Schaer, and Po-Yu Chuang. Overview of the trec 2016 open search track. In *Text REtrieval Conference (TREC), Gaithersburg, Maryland, USA*, 2016. *Cité page 95*

Michele Banko, Michael J Cafarella, Stephen Soderland, Matt Broadhead, and Oren Etzioni. Open information extraction from the web. In *Proc. of IJCAI 2007*, pages 2670–2676, 2007. *Cité page 42*

Massimo Bartoletti, Stefano Lande, and Alessandro Massa. Faderank : An incremental algorithm for ranking twitter users. In Wojciech Cellary, Mohamed F. Mokbel, Jianmin Wang, Hua Wang, Rui Zhou, and Yanchun Zhang, editors, *Web Information Systems Engineering – WISE 2016*, pages 55–69. Springer International Publishing, 2016. *Cité page 29*

Gaurav Baruah, Rakesh Guttikonda, Adam Roegiest, and Olga Vechtomova. University of Xaterloo at the TREC 2013 temporal summarization track. In *Text REtrieval Conference (TREC), Gaithersburg, Maryland, USA*, 2013. *Cité page 59*

Senjuti Basu Roy, Sihem Amer-Yahia, Ashish Chawla, Gautam Das, and Cong Yu. Constructing and exploring composite items. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data*, pages 843–854. ACM, 2010. *2 citations pages 54 et 55*

Mikhail Bautin and Steven Skiena. Concordance-based entity-oriented search. *Web Intelligence and Agent Systems : An International Journal*, 7(4) :303–319, 2009. https://doi.org/10.1109/WI.2007.84. *2 citations pages 40 et 41*

Kedar Bellare, Partha Pratim Talukdar, Giridhar Kumaran, O Pereira, Mark Liberman, Andrew Mccallum, and Mark Dredze. Lightly-supervised attribute extraction for web search. In *Proc. of Machine Learning for Web Search Workshop, NIPS 2007*, 2007. *2 citations pages 41 et 46*

Mohamed Ben Aouicha, M. Tmar, and Mohand Boughanem. Flexible document-query matching based on a probabilistic content and structure score combination. In *Symposium on Applied Computing (SAC), Sierre, Switzerland*. ACM, mars 2010. *Cité page 14*

Lamjed Ben Jabeur, Firas Damak, Lynda Tamine, Karen Pinel-Sauvagnat, Guillaume Cabanac, and Mohand Boughanem. IRIT at TREC Microblog 2012 : Adhoc Task. In Ellen M. Voorhees and Lori P. Buckland, editors, *Text REtrieval Conference (TREC), Gaithersburg, USA*, http://www-nlpir.nist.gov/, novembre 2012a. National Institute of Standards and Technology (NIST). http://trec.nist.gov/pubs/trec21/papers/IRIT.microblog.final.pdf. *4 citations pages 34, 37, 71, et V*

Lamjed Ben Jabeur, Lynda Tamine, and Mohand Boughanem. Active microbloggers : Identifying influencers, leaders and discussers in microblogging networks. In *String Processing and Information Retrieval*, volume 7608 of *Lecture Notes in Computer Science*, pages 111–117. Springer Berlin Heidelberg, 2012b. *Cité page 29*

Lamjed Ben Jabeur, Firas Damak, Lynda Tamine, Guillaume Cabanac, Karen Pinel-Sauvagnat, and Mohand Boughanem. IRIT at TREC Microblog Track 2013. In Ellen M. Voorhees, editor, *Text REtrieval Conference (TREC), Gaithersburg, USA*, http://www-nlpir.nist.gov/, novembre 2013. National Institute of Standards and Technology (NIST). http://trec.nist.gov/pubs/trec22/papers/IRIT-microblog.pdf. *5 citations pages 30, 31, 37, 71, et V*

Scott Boag, Don Chamberlin, Mary F. Fernández, Daniela Florescu, Jonathan Robie, and Jérôme Siméon. XQuery 1.0 : An XML query language. Technical report, World Wide Web Consortium (W3C), W3C Recommandation, december 2010. http://www.w3.org/TR/2010/REC-xquery-20101214/. *Cité page 10*

Ludovic Bonnefoy, Vincent Bouvier, and Patrice Bellot. 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*, pages 769–772. ACM, 2013. *Cité page 43*

Horatiu Bota, Ke Zhou, Joemon M Jose, and Mounia Lalmas. Composite retrieval of heterogeneous web search. In *Proceedings of the 23rd international conference on World wide web*, pages 119–130. ACM, 2014. *Cité page 55*

Julien Boujot. Utilisation des liens pour la recherche dans les documents structurés : XMLRank . Rapport de master, IRIT, Université Paul Sabatier, Toulouse, juin 2007. *Cité page 26*

Vincent Bouvier and Patrice Bellot. Filtering entity centric documents using numerics and temporals features within rf classifier. In *Text REtrieval Conference (TREC), Gaithersburg, Maryland, USA*, 2013. *Cité page 43*

Luca Busin and Stefano Mizzaro. Axiometrics : An axiomatic approach to information retrieval effectiveness metrics. In *Proceedings of the 2013 Conference on the Theory of Information Retrieval*, ICTIR '13, pages 8 :22–8 :29. ACM, 2013. ISBN 978-1-4503-2107-5. *Cité page 97*

Michael J. Cafarella, Michele Banko, and Oren Etzioni. Relational web search. Technical report, University of Washington, 2006. *2 citations pages 55 et 56*

Michael J. Cafarella, Alon Y. Halevy, Yang Zhang, Daisy Zhe Wang, and Eugene Wu. Uncovering the relational web. In *Proceedings of WebDB*, 2008. *2 citations pages 43 et 44*

Michael J. Cafarella, Alon Y. Halevy, and Nodira Khoussainova. Data integration for the relational web. *PVLDB*, 2(1) :1090–1101, 2009. https://doi.org/10.14778/1687627.1687750. *Cité page 41*

Jamie Callan. Distributed information retrieval. In W. Bruce Croft, editor, *Advances in Information Retrieval*, pages 235–266. Kluwer Academic Publishers, 2000. *2 citations pages 3 et 54*

Ricardo Campos, Gaël Dias, Alípio M Jorge, and Adam Jatowt. Survey of temporal information retrieval and related applications. *ACM Computing Surveys (CSUR)*, 47(2) :15, 2015. *Cité page 41*

Linda Cappellato, Nicola Ferro, Gareth J.F. Jones, Jaap Kamps, Josiane Mothe, Karen Pinel-Sauvagnat, Eric San Juan, and Jacques Savoy. Report on CLEF 2015 : Experimental IR Meets Multilinguality, Multimodality, and Interaction. *SIGIR Forum*, 49(2) :(47–56, décembre 2015. http://doi.org/10.1145/2888422.2888428. *Cité page 85*

Claudio Carpineto, Stanislaw Osiński, Giovanni Romano, and Dawid Weiss. A survey of web clustering engines. *ACM Comput. Surv.*, 41(3) :1–38, 2009. https://doi.org/10.1145/1541880.1541884. *Cité page 72*

Ben Carterette, Virgil Pavlu, Evangelos Kanoulas, Javed A. Aslam, and James Allan. Evaluation over thousands of queries. In *SIGIR '08 : Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 651–658. ACM, 2008. *Cité page 79*

Pat Case, Michael Dyck, Mary Holstege, Sihem Amer-Yahia, Chavdar Botev, Stephen Buxton, Jochen Doerre, Jim Melton, Michael Rys, and Jayavel Shanmugasundaram. XQuery 1.0 : An XML query language. Technical report, World Wide Web Consortium (W3C), W3C Recommandation, march 2011. http://www.w3.org/TR/2011/REC-xpath-full-text-10-20110317/. *Cité page 10*

Angel X Chang and Christopher D Manning. Sutime : A library for recognizing and normalizing time expressions. In *Lrec*, volume 2012, pages 3735–3740, 2012. *Cité page 50*

Chia-Hui Chang, M. Kayed, M. R. Girgis, and K. F. Shaalan. A survey of web information extraction systems. *IEEE Transactions on Knowledge and Data Engineering*, 18(10) :1411–1428, Oct 2006. http://doi.org/10.1109/TKDE.2006.152. *Cité page 44*

Abdelhamid Chellal, Lamjed Ben Jabeur, Laure Soulier, Bilel Moulahi, Thomas Palmer, Mohand Boughanem, Karen Pinel-Sauvagnat, Lynda Tamine, and Gilles Hubert. IRIT at TREC Microblog 2015 . In *Text REtrieval Conference (TREC), Gaithersburg, Maryland USA*, http://www.nist.org, novembre 2015. National Institute of Standards and Technology (NIST). http://trec.nist.gov/act_part/conference/papers/IRIT-MB.pdf. *4 citations pages 36, 37, 71, et V*

Hsin-Hsi Chen, Shih-Chung Tsai, and Jin-He Tsai. Mining tables from large scale html texts. In *Proceedings of the 18th Conference on Computational Linguistics - Volume 1*, COLING '00, pages 166–172. Association for Computational Linguistics, 2000. *Cité page 44*

Lei Chen, Hainan Zhang, Siying Li, Zhiyuan Ji, Qian Liu, Yue Liu, Dayong Wu, and Xueqi Cheng. Ictnet at temporal summarization track trec 2014. In *Text REtrieval Conference (TREC), Gaithersburg, Maryland, USA*, 2014. *Cité page 60*

Peter Pin-Shan Chen. The entity-relationship model-toward a unified view of data. *ACM Trans. Database Syst.*, 1(1) :9–36, 1976. http://doi.acm.org/10.1145/320434.320440. *2 citations pages 2 et 40*

Weimin Chen. More efficient algorithm for ordered tree inclusion. *Journal of Algorithms*, 26 :370–385, 1998. https://doi.org/10.1006/jagm.1997.0899. *Cité page 14*

Fuxing Cheng, Xin Zhang, Ben He, Tiejian Luo, and Wenjie Wang. A survey of learning to rank for real-time twitter search. In *Pervasive computing and the networked world*, pages 150–164. Springer, 2012. *Cité page 37*

Cyril W. Cleverdon. The Cranfield tests on index languages devices. In *Aslib Proceedings, volume 19, pages 173-192*, 1967. *2 citations pages 71 et 95*

Doron Cohen, Einat Amitay, and David Carmel. Lucene and juru at trec 2007 : 1-million queries track. In *Text REtrieval Conference (TREC), Gaithersburg, Maryland, USA*, 2007. *Cité page 32*

Daniel Crabtree, Xiaoying Gao, and Peter Andreae. Standardized evaluation method for web clustering results. In *IEEE/WIC/ACM International Conference on Web Intelligence*, pages 280–283, 2005. *Cité page 74*

Valter Crescenzi, Giansalvatore Mecca, and Paolo Merialdo. Roadrunner : Towards automatic data extraction from large web sites. In *Proc. of VLDB 2001*, pages 109–118, 2001. *Cité page 41*

Theodore Dalamagas, Tao Cheng, Klaas-jan Winkel, and Timos Sellis. A methodology for clustering xml documents by structure. *Information Systems*, 31 :187–228, 2006. https://doi.org/10.1016/j.is.2004.11.009. *Cité page 15*

Firas Damak. 'Étude de l'impact de l'agrégation des recherches verticales sur la qualités des SRI. Rapport de master, IRIT, Université Paul Sabatier, Toulouse, Juin 2010. *Cité page 68*

Firas Damak. *Étude des facteurs de pertinence dans la recherche de microblogs*. Thèse de doctorat, Université Paul Sabatier, Toulouse, France, juillet 2014. https://tel.archives-ouvertes.fr/tel-01074732, Soutenance le 15/07/2014. *9 citations pages iv, 28, 29, 30, 31, 35, 36, 38, et 89*

Firas Damak, Lamjed Ben Jabeur, Guillaume Cabanac, Karen Pinel-Sauvagnat, Lynda Tamine, and Mohand Boughanem. IRIT at TREC Microblog 2011. In Voorhees Ellen M. and Buckland Lori P., editors, *Text REtrieval Conference (TREC), Gaithersburg, USA*, http://www-nlpir.nist.gov/, novembre 2011. National Institute of Standards and Technology (NIST). http://trec.nist.gov/pubs/trec20/papers/IRIT_SIG.microblog.update.pdf. *4 citations pages 32, 37, 71, et V*

Firas Damak, Karen Pinel-Sauvagnat, and Guillaume Cabanac. Recherche de microblogs : quels critères pour raffiner les résultats des moteurs usuels de RI ? In *Conférence francophone en Recherche d'Information et Applications (CORIA), Bordeaux*, pages 317–328. LABRI, mars 2012. doi:10.24348/coria.2012.317. *3 citations pages 34, 38, et V*

Firas Damak, Karen Pinel-Sauvagnat, Guillaume Cabanac, and Mohand Boughanem. Effectiveness of State-of-the-art Features for Microblog Search . In *ACM Symposium on Applied Computing (SAC), Coimbra, Portugal*, pages 914–919. ACM, mars 2013. http://dx.doi.org/10.1145/2480362.2480537. *6 citations pages vi, 31, 32, 34, 36, et V*

Ernesto Damiani, Barbara Oliboni, and Letizia Tanca. Fuzzy techniques for xml data smushing. In *Proceedings of the International Conference, 7th Fuzzy Days on Computational Intelligence, Theory and Applications*, pages 637–652. Springer-Verlag, 2001. *Cité page 14*

Ernestro Damiani, Letizia Tanca, and Fontana Arcelli. Fuzzy xml queries via context-based choice of aggregation. *Kybernetika*, 36 :635–655, 2000. *Cité page 14*

Ritendra Datta, Jia Li, and James Z. Wang. Content-based image retrieval : approaches and trends of the new age. In *ACM SIGMM international workshop on Multimedia information retrieval*, pages 253–262, 2005. *Cité page 76*

Gianluca Demartini, Tereza Iofciu, and Arjen P De Vries. Overview of the INEX 2009 entity ranking track. In *Focused Retrieval and Evaluation*, pages 254–264. Springer, 2009. *2 citations pages 11 et 97*

Ludovic Denoyer and Patrick Gallinari. Overview of the INEX 2008 XML mining track. In *Advances in Focused Retrieval*, pages 401–411. Springer, 2008. *Cité page 11*

Fernando Diaz. Integration of news content into web results. In *WSDM*, pages 182–191, 2009. *Cité page 78*

Laura Dietz and Jeffrey Dalton. Umass at trec 2013 knowledge base acceleration track : Bi-directional entity linking and time-aware evaluation. In *Text REtrieval Conference (TREC), Gaithersburg, Maryland, USA*, 2013. *Cité page 43*

Laura Dietz, Alexander Kotov, and Edgar Meij. Utilizing knowledge bases in text-centric information retrieval. In *Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval*, pages 5–5. ACM, 2016. *Cité page 41*

Yajuan Duan, Long Jiang, Tao Qin, Ming Zhou, and Heung-Yeung Shum. An empirical study on learning to rank of tweets. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 295–303, 2010. *3 citations pages 32, 33, et 37*

Elizabeth Dubois and Devin Gaffney. The multiple facets of influence : Identifying political influentials and opinion leaders on twitter. *American Behavioral Scientist*, 58(10) :1260–1277, 2014. https://doi.org/10.1177%2F0002764214527088. *Cité page 29*

Serge Dulucq and Helene Touzet. Analysis of tree edit distance algorithms. In *Proceedings of the 14th annual symposium of combinatorial pattern matching*, pages 83–95, 2003. *Cité page 15*

Hazem Elmeleegy, Jayant Madhavan, and Alon Y. Halevy. Harvesting relational tables from lists on the web. *The VLDB Journal*, 2(1) :209–226, 2009. https://doi.org/10.1007/s00778-011-0223-0. *Cité page 58*

Oren Etzioni, Michael Cafarella, Doug Downey, Ana-Maria Popescu, Tal Shaked, Stephen Soderland, Daniel S. Weld, and Alexander Yates. Unsupervised named-entity extraction from the web : an experimental study. *Artif. Intell.*, 165(1) :91–134, 2005. http://dx.doi.org/10.1016/j.artint.2005.03.001. *Cité page 41*

Oren Etzioni, Anthony Fader, Janara Christensen, Stephen Soderland, and Mausam. Open information extraction : The second generation. In *Proc. of IJCAI 2011, Barcelona, Spain*, pages 3–10, 2011. *Cité page 42*

Anthony Fader, Stephen Soderland, and Oren Etzioni. Identifying relations for open information extraction. In *Proceedings of the conference on empirical methods in natural language processing*, pages 1535–1545. Association for Computational Linguistics, 2011. *Cité page 42*

Robert W. Floyd. Algorithm 97 : Shortest path. *Commun. ACM*, 5 :345, June 1962. https://doi.org/10.1145/367766.368168. *Cité page 16*

John R. Frank, Max Kleiman-Weiner, Daniel A. Roberts, Feng Niu, Ce Zhang, Christopher Ré, and Ian Soboroff. Building an Entity-Centric stream filtering test collection for TREC 2012. In *Text REtrieval Conference (TREC), Gaithersburg, Maryland, USA*, 2012. *Cité page 42*

John R Frank, Steven J Bauer, Max Kleiman-Weiner, Daniel A Roberts, Nilesh Tripuraneni, Ce Zhang, Christopher Ré, Ellen Voorhees, and Ian Soboroff. Evaluating stream filtering for entity profile updates for TREC 2013 (KBA Track Overview). In *Text REtrieval Conference (TREC), Gaithersburg, Maryland, USA*, 2013. *4 citations pages 51, 59, 97, et V*

John R Frank, Max Kleiman-Weiner, Daniel A Roberts, Ellen M Voorhees, and Ian Soboroff. Evaluating stream filtering for entity profile updates in trec 2012, 2013, and 2014. In *Text REtrieval Conference (TREC), Gaithersburg, Maryland, USA*, 2014. *3 citations pages 42, 51, et V*

Norbert Fuhr. Some common mistakes in IR evaluation, and how they can be avoided. *SIGIR Forum*, 51(3) :32–41, 2017. https://doi.org/10.1145/3190580.3190586. *Cité page 95*

Norbert Fuhr and Mounia Lalmas. Report on the inex 2003 workshop. *SIGIR Forum*, 38 :46–51, 2004. https://doi.org/10.1145/986278.986287. *Cité page III*

Salvatore Gaglio, Giuseppe Lo Re, and Marco Morana. Real-time detection of twitter social events from the user's perspective. In *Communications (ICC), 2015 IEEE International Conference on*, pages 1207–1212. IEEE, 2015. *Cité page 37*

Jeremy Goecks. Nuggetmine : Intelligent groupware for opportunistically sharing information nuggets. In *Proc. of IUI '02*, pages 87–94, 2002. *Cité page 2*

Jade Goldstein, Vibhu Mittal, Jaime Carbonell, and Mark Kantrowitz. Multi-document summarization by sentence extraction. In *Proceedings of the 2000 NAACL-ANLPWorkshop on Automatic Summarization - Volume 4*, pages 40–48, 2000. http://dx.doi.org/10.3115/1117575.1117580. *2 citations pages 3 et 54*

Zhiguo Gong, Hou Leong Hou, and Chan Wa Cheang. Web image indexing by using associated texts. *Knowledge and Information Systems*, pages 243–264, 2006. https://doi.org/10.1007/s10115-005-0231-8. *Cité page 21*

Norbert Govert, Mohamed Abolhassani, Norbert Fuhr, and Kai Grossjohann. Content-oriented XML retrieval with HyReX. In *Proceedings INEX 2002, Dagstuhl, Germany*, 2002. *Cité page 10*

Michael Grubinger, Paul Clough, Allan Hanbury, and Henning Müller. Overview of the ImageCLEFphoto 2007 Photographic Retrieval Task. In Carol Peters, Valentin Jijkoun, Thomas Mandl, Henning Müller, Douglas W. Oard, Anselmo Peñas, Vivien Petras, and Diana Santos, editors, *Advances in Multilingual and Multimodal Information Retrieval : 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007, Budapest, Hungary, September 19-21, 2007, Revised Selected Papers*, pages 433–444. Springer Berlin Heidelberg, 2008. *Cité page IV*

Kevin Haas, Peter Mika, Paul Tarjan, and Roi Blanco. Enhanced results for web search. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 725–734. ACM, 2011. http://doi.acm.org/10.1145/2009916.2010014. *Cité page 2*

Maria Halkidi, Yannis Batistakis, and Michalis Vazirgiannis. On clustering validation techniques. *J. Intell. Inf. Syst.*, 17(2-3) :107–145, 2001. https://doi.org/10.1023/A:1012801612483. *Cité page 72*

Mark A. Hall and Geoffrey Holmes. Benchmarking attribute selection techniques for discrete class data mining. *IEEE Trans. on Knowl. and Data Eng.*, 15(6) :1437–1447, 2003. http://doi.org/10.1109/TKDE.2003.1245283. *Cité page 34*

Vassilis Harmandas, Mark Sanderson, and Mark Dunlop. Image retrieval by hypertext links. In *The 20th annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR'97*, pages 296–303, 1997. *Cité page 21*

Qi He, Kuiyu Chang, Ee-Peng Lim, and Jun Zhang. Bursty feature representation for clustering text streams. In *Proceedings of the 2007 SIAM International Conference on Data Mining*, pages 491–496. SIAM, 2007. *Cité page 43*

Marti A Hearst. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th conference on Computational linguistics-Volume 2*, pages 539–545. Association for Computational Linguistics, 1992. *Cité page 47*

Marti A Hearst. Automated discovery of wordnet relations. *WordNet : an electronic lexical database*, pages 131–153, 1998. *Cité page 47*

Sascha Hennig and Michael Wurst. Incremental clustering of newsgroup articles. In *Advances in Applied Artificial Intelligence : 19th International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, IEA/AIE 2006, Annecy, France, June 27-30, 2006. Proceedings*. Springer Berlin Heidelberg, 2006. https://doi.org/10.1007/11779568_37. *2 citations pages 4 et 55*

Djoerd Hiemstra. Statistical language models for intelligent xml retrieval. In Burkhard Stiller, Georg Carle, Martin Karsten, and Peter Reichl, editors, *Group Communications and Charges. Technology and Business Models*, volume 2818 of *Lecture Notes in Computer Science*, pages 107–118. Springer Berlin / Heidelberg, 2003. *Cité page 16*

Lobna Hlaoua. *Reformulation de Requêtes par Réinjection de Pertinence dans les documents Semi-Structurés.* Thèse de doctorat, Université Paul Sabatier, Toulouse, France, décembre 2007. http://thesesups.ups-tlse.fr/114/1/Hlaoua_Lobna.pdf, Soutenance le 14/12/2007). *2 citations pages 26 et 89*

Lobna Hlaoua and Karen Pinel-Sauvagnat. Structure-Oriented Relevance Feedback in XML Retrieval. In Vicente P. Guerrero-Note, editor, *International Conference on Multidisciplinary Sciences & Technologies (InSciT), Merida, Espagne*, pages 99–103. Open Institute of Knowledge, octobre 2006. *2 citations pages 19 et III*

Lobna Hlaoua, Mohand Boughanem, and Karen Sauvagnat. Réinjection de structures pour la reformulation de requêtes en RI structurée . In *Congrès Informatique des Organisations et Systèmes d'Information et de Décision (INFORSID), Hammamet*, pages 435–450. INFORSID (actes électroniques), 2006a. http://www.irit.fr/~Karen.Pinel-Sauvagnat/fichiers/INFORSID2006.pdf. *2 citations pages 19 et III*

Lobna Hlaoua, Karen Pinel-Sauvagnat, and Mohand Boughanem. A structure-oriented relevance feedback method for XML retrieval. In *Conference on Information and Knowledge Management (CIKM), Arlington, Virginia, USA.*, pages 780–781. ACM, novembre 2006b. http://dx.doi.org/10.1145/1183614.1183727. *3 citations pages 19, 25, et III*

Lobna Hlaoua, Mohand Boughanem, and Karen Pinel-Sauvagnat. Using a Content-and-Structure Oriented Method for Relevance Feedback in XML Retrieval. In *Large-Scale Semantic Access to Content (Text, Image, Video and Sound) (RIAO), Pittsburgh (PA) États-Unis*. Centre de hautes études internationales d'Informatique Documentaire (C.I.D.), juin 2007a. *2 citations pages 20 et III*

Lobna Hlaoua, Mohand Boughanem, and Karen Pinel-Sauvagnat. Combination of Evidences in Relevance Feedback for XML Retrieval. In *Conference on Information and Knowledge Management (CIKM), Lisbonne, Portugal*, pages 893–896. ACM Press, novembre 2007b. http://dx.doi.org/10.1145/1321440.1321569. *3 citations pages 20, 25, et III*

Lobna Hlaoua, Mohand Boughanem, and Karen Pinel-Sauvagnat. Combinaison des caractéristiques des termes pour l'extension de requêtes en recherche d'information dans les documents semi-structurés. In *Conférence francophone en Recherche d'Information et Applications (CORIA), Saint Etienne France*, pages 77–92, http ://portail.univ-st-etienne.fr/, mars 2007c. Université de Saint-Etienne. http://doi.org/doi:10.24348/coria.2007.77. *2 citations pages 19 et III*

Lobna Hlaoua, Karen Pinel-Sauvagnat, and Mohand Boughanem. Relevance Feedback for XML Retrieval : using structure and content to expand queries . In Colette Rolland, Oscar Pastor, and Jean-Louis Cavarero, editors, *International Conference on Research Challenge in Information Science (RCIS), Ouarzazate- Maroc*, pages 195–202. EMSI - Ecole Marocaine des Sciences de l'Ingénieur, avril 2007d. *2 citations pages 19 et III*

Lobna Hlaoua, Mouna Torjmen, Karen Pinel-Sauvagnat, and Mohand Boughanem. XFIRM at INEX 2006. Ad-hoc, Relevance Feedback and MultiMedia tracks. In Norbert Fuhr, Mounia Lalmas, and Andrew Trotman, editors, *International Workshop of the Initiative for the Evaluation of XML Retrieval (INEX), Dagstuhl, Allemagne*, volume LNCS 4518, pages 373–386. Springer, mars 2007e. https://doi.org/10.1007/978-3-540-73888-6_36. *7 citations pages 18, 19, 23, 25, 71, III, et IV*

Lobna Hlaoua, Karen Pinel-Sauvagnat, and Mohand Boughanem. Relevance Feedback Revisited : Dealing with Content and Structure in XML Documents. *International Journal on Digital Libraries*, 11(1) :1–24, Mars 2010. http://dx.doi.org/10.1007/s00799-010-0061-5. *5 citations pages 18, 19, 20, 25, et III*

Liangjie Hong, Byron Dom, Siva Gurumurthy, and Kostas Tsioutsiouliklis. A time-dependent topic model for multiple text streams. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, pages 832–840. ACM, 2011. *Cité page 4*

Gilles Hubert, Guillaume Cabanac, Karen Pinel-Sauvagnat, Damien Palacio, and Christian Sallaberry. IRIT, GeoComp, and LIUPPA at the TREC 2013 Contextual Suggestion Track. In Ellen M. Voorhees, editor, *Text REtrieval Conference (TREC), Gaithersburg, USA*, http://www-nlpir.nist.gov/, novembre 2013. National Institute of Standards and Technology (NIST). http://trec.nist.gov/pubs/trec22/papers/IRIT-context.pdf. *Cité page 91*

Gilles Hubert, José Moreno, Karen Pinel-Sauvagnat, and Yoann Pitarch. Everything You Always Wanted to Know About TREC RTS* (*But Were Afraid to Ask). Diffusion scientifique, décembre 2017a. https://arxiv.org/abs/1712.04671. *6 citations pages 37, 82, 83, 84, 95, et V*

Gilles Hubert, José Moreno, Karen Pinel-Sauvagnat, and Yoann Pitarch. Some thoughts from IRIT about the scenario A of the TREC RTS 2016 and 2017 tracks. In *Text REtrieval Conference (TREC), Gaithersburg, Maryland, USA*, novembre 2017b. http://trec.nist.gov/pubs/trec26/papers/IRIT2-RT.pdf. *7 citations pages 37, 71, 82, 83, 84, 95, et V*

Gilles Hubert, Yoann Pitarch, Karen Pinel-Sauvagnat, Ronan Tournier, and Léa Laporte. TournaRank : When Retrieval Becomes Document Competition. *Information Processing & Management*, 54(2) :252–272, mars 2018a. https://doi.org/10.1016/j.ipm.2017.11.006. *2 citations pages 93 et 94*

Gilles Hubert, Yoann Pitarch, Karen Pinel-Sauvagnat, Ronan Tournier, and Léa Laporte. TournaRank : Quand la Recherche d'Information devient un tournoi entre documents. In *Conférence francophone en Recherche d'Information et Applications (CORIA), Rennes*, mai 2018b. *2 citations pages 93 et 94*

Diego Ingaramo, David Pinto, Paolo Rosso, and Marcelo Errecalde. Evaluation of internal validity measures in short-text corpora. In *Computational linguistics and intelligent text processing*, pages 555–567. Springer-Verlag, 2008. *Cité page 72*

Bernard J Jansen, Amanda Spink, and Tefko Saracevic. Real life, real users, and real needs : a study and analysis of user queries on the web. *Information Processing & Management*, 36(2) :207 – 227, 2000. https://doi.org/10.1016/S0306-4573(99)00056-4. *Cité page 2*

Nicholas Jardine and Cornelis Joost van Rijsbergen. The use of hierarchic clustering in information retrieval. *Inform. Stor. Retr.*, 7(5) :217–240, 1971. https://doi.org/10.1016/0020-0271(71)90051-9. *2 citations pages 72 et 73*

Fred Jelinek and Robert L. Mercer. Interpolated estimation of Markov source parameters from sparse data. In *In Proceedings of the Workshop on Pattern Recognition in Practice*, pages 381–397, 1980. *Cité page 18*

Heng Ji, Joel Nothman, Ben Hachey, et al. Overview of tac-kbp2014 entity discovery and linking tasks. In *Proc. Text Analysis Conference (TAC2014)*, pages 1333–1339, 2014. *Cité page 97*

Jingtian Jiang, Chin-Yew Lin, and Yong Rui. MSR KMG at TREC 2014 KBA track vital filtering task. In *Text REtrieval Conference (TREC), Gaithersburg, Maryland, USA*, 2014. *Cité page 43*

Tao Jiang, Lusheng Wang, and Kaizhong Zhang. Alignment of trees - an alternative to tree edit. *Theoritical Computer Science*, pages 137–148, 1995. https://doi.org/10.1016/0304-3975(95)80029-9. *Cité page 14*

Hervé Jégou, Matthijs Douze, and Cordelia Schmid. Hamming embedding and weak geometric consistency for large scale image search. In *Proceedings of ECCV*, 2008. *Cité page 77*

Jaap Kamps, Maarten de Rijke, and Borkur Sigurbjornsson. Length normalization in XML retrieval. In *Proceedings of SIGIR 2004, Sheffield, England*, pages 80–87, 2004. *Cité page 10*

Rianne Kaptein and Maarten Marx. Focused retrieval and result aggregation with political data. *Inf. Retr.*, 13 :412–433, October 2010. https://doi.org/10.1007/s10791-010-9130-z. *Cité page 55*

Gabriella Kazai. Initiative for the evaluation of xml retrieval. In Ling Liu and M. Tamer Özsu, editors, *Encyclopedia of Database Systems*, pages 1531–1537. Springer US, 2009. ISBN 978-0-387-39940-9. doi : 10.1007/978-0-387-39940-9_151. http://dx.doi.org/10.1007/978-0-387-39940-9_151. *Cité page 10*

Gabriella Kazai and Mounia Lalmas. extended cumulated gain measures for the evaluation of content-oriented XML retrieval. *ACM Transactions on Information Systems (TOIS)*, 24(4) :503–542, 2006. https://doi.org/10.1145/1185877.1185883. *Cité page 10*

Gabriella Kazai, Mounia Lalmas, and Arjen P. de Vries. The overlap problem in content-oriented XML retrieval evaluation. In *Proceedings of SIGIR 2004, Sheffield, England*, pages 72–79, July 2004. *Cité page 10*

Lyndon S. Kennedy and Mor Naaman. Generating diverse and representative image search results for landmarks. In *Proceedings of the 17th International Conference on World Wide Web*, WWW '08, pages 297–306, 2008. *2 citations pages 4 et 55*

Farhan Hassan Khan, Saba Bashir, and Usman Qamar. Tom : Twitter opinion mining framework using hybrid classification scheme. *Decision Support Systems*, 57 :245 – 257, 2014. https://doi.org/10.1016/j.dss.2013.09.004. *Cité page 29*

Hyun Duk Kim, Kavita Ganesan, Parikshit Sondhi, and Cengxiang Zhai. Comprehensive review of opinion summarization. Rapport technique, University of Illinois at Urbana-Champaign, 2011. *Cité page 55*

Philip N. Klein. Computing the edit-distance between unrooted ordered trees. In *Proceedings of the 6th Annual European Symposium on Algorithms*, ESA '98, pages 91–102. Springer-Verlag, 1998. *Cité page 15*

Zhigang Kong and Mounia Lalmas. Using XML logical structure to retrieve (multimedia) objects. In *European Conference on Digital Libraries, ECDL'07*, pages 100, 111, 2007a. *Cité page 23*

Zhigang Kong and Mounia Lalmas. Combining multiple sources of evidence in XML multimedia documents : An inference network incorporating element language models. In *29th European Conference on Information Retrieval (Poster), ECIR'07*, pages 716–719, 2007b. *Cité page 23*

Arlind Kopliku. *Approaches to implement and evaluate aggregated search.* Thèse de doctorat, Université Paul Sabatier, Toulouse, France, décembre 2011. http://thesesups.ups-tlse.fr/1535/, Soutenance le 07/12/2011. *3 citations pages 52, 68, et 90*

Arlind Kopliku, Mohand Boughanem, and Karen Pinel-Sauvagnat. Querying by examples. In *Conférence francophone en Recherche d'Information et Applications (CORIA), Sousse, Tunisie*, pages 407–408. Association Francophone de Recherche d'Information et Applications (ARIA), mars 2010. http://doi.org/doi:10.24348/coria.2010.407. *Cité page 47*

Arlind Kopliku, Mohand Boughanem, and Karen Pinel-Sauvagnat. Towards a framework for attribute retrieval . In *Conference on Information and Knowledge Management (CIKM), Glasgow, UK*, pages 515–524. ACM, octobre 2011a. https://doi.org/10.1145/2063576.2063654. *9 citations pages iv, v, 41, 43, 44, 45, 46, 51, et 61*

Arlind Kopliku, Mohand Boughanem, and Karen Pinel-Sauvagnat. Mining the Web for lists of Named Entities. In *Conférence francophone en Recherche d'Information et Applications (CORIA), Avignon*, pages 113–120. Association Francophone de Recherche d'Information et Applications (ARIA), mars 2011b. http://doi.org/doi:10.24348/coria.2011.113. *Cité page 47*

Arlind Kopliku, Firas Damak, Karen Pinel-Sauvagnat, and Mohand Boughanem. Interest and Evaluation of Aggregated Search . In *IEEE/WIC/ACM International Conference on Web Intelligence, Lyon*. ACM, août 2011c. https://doi.org/110.1109/WI-IAT.2011.99. *5 citations pages v, 4, 78, 79, et 80*

Arlind Kopliku, Karen Pinel-Sauvagnat, and Mohand Boughanem. Retrieving attributes using Web tables. In *Joint Conference on Digital Libraries (JCDL) (JCDL), Ottawa*, pages 397–398. ACM, juin 2011d. http://dx.doi.org/10.1145/1998076.1998153. *2 citations pages 43 et 51*

Arlind Kopliku, Karen Pinel-Sauvagnat, and Mohand Boughanem. Attribute Retrieval from Relational Web tables . In *Symposium on String Processing and Information Retrieval (SPIRE), Pisa, Italy*, pages 117–128. Springer, octobre 2011e. https://doi.org/10.1007/978-3-642-24583-1_12. *4 citations pages 4, 43, 44, et 51*

Arlind Kopliku, Karen Pinel-Sauvagnat, and Mohand Boughanem. Aggregated search : a new information retrieval paradigm. *ACM Computing Surveys*, 46(3) :1–31, janvier 2014. http://doi.acm.org/10.1145/2523817. *5 citations pages iv, 2, 4, 41, et 54*

Sherry Koshman, Amanda Spink, and Bernard J. Jansen. Web searching on the Vivisimo search engine. *JASIST*, 57(14) :1875–1887, 2006. https://doi.org/10.1002/asi.v57:14. *Cité page 72*

Ines Krichen. Extraction, sélection et agrégation d'information à partir d'une base de connaissance. Rapport de master, IRIT, Université Paul Sabatier, Toulouse, Juin 2010. *Cité page 68*

Ines Krichen, Arlind Kopliku, Karen Pinel-Sauvagnat, and Mohand Boughanem. Une approche de recherche d'attributs pertinents pour l'agrégation d'information . In *INFormatique des Organisations et Systemes d'Information et de Decision (INFORSID), Lille*, pages 385–400, http ://inforsid.irit.fr/, mai 2011. Association INFORSID. ftp://ftp.irit.fr/IRIT/SIG/Recherche_d_attributs_IK_AK_KPS_MB_Inforsid2011.pdf. *Cité page 63*

Ines Krichen, Arlind Kopliku, Karen Pinel-Sauvagnat, and Mohand Boughanem. Une approche de recherche d'attributs pertinents pour l'agrégation d'information. *Document numérique*, 15(1) : 9–32, juin 2012. https://www.cairn.info/revue-document-numerique-2012-1-page-9.htm. *3 citations pages v, 61, et 63*

Cyril Laitang. *Impact de la structure des documents XML sur le processus d'appariement dans le contexte de la Recherche d'Information Semi-structurée*. Thèse de doctorat, Université Paul Sabatier, Toulouse, France, Juillet 2013. http://thesesups.ups-tlse.fr/2091/1/2013TOU30099.pdf, (Soutenance le 12/07/2013). *4 citations pages 13, 16, 26, et 89*

Cyril Laitang and Karen Pinel-Sauvagnat. Utilisation de la théorie des graphes et de la distance d'édition pour la recherche d'information sur documents XML . In *Conférence francophone en Recherche d'Information et Applications (CORIA), Avignon*, pages 349–364. Association Francophone de Recherche d'Information et Applications (ARIA), mars 2011. http://doi.org/doi:10.24348/coria.2011.349. *3 citations pages 12, 14, et IV*

Cyril Laitang, Mohand Boughanem, and Karen Pinel-Sauvagnat. XML Information Retrieval through Tree Edit Distance and Structural Summaries. In *Asia Information Retrieval Society Conference (AIRS), Dubai, United Atab Emirates*, pages 73–83. Springer, décembre 2011. http://dx.doi.org/10.1007/978-3-642-25631-8_7. *3 citations pages 12, 15, et IV*

Cyril Laitang, Karen Pinel-Sauvagnat, and Mohand Boughanem. Edit Distance for XML Information Retrieval : Some Experiments on the Datacentric Track of INEX 2011. In *Focused Retrieval of Content and Structure - INEX (Initiative for the Evaluation of XML Retrieval), Imsbach*, Lecture note in computer science, pages 138–145. Springer, avril 2012a. https://doi.org/10.1007/978-3-642-35734-3_11. *4 citations pages 12, 25, 71, et IV*

Cyril Laitang, Karen Pinel-Sauvagnat, and Mohand Boughanem. Coûts de distance d'édition pour la Recherche d'Information XML . In *Conférence francophone en Recherche d'Information et Applications (CORIA), Bordeaux*, pages 357–372, mars 2012b. http://doi.org/doi:10.24348/coria.2012.357. *4 citations pages 12, 13, 16, et IV*

Cyril Laitang, Karen Pinel-Sauvagnat, and Mohand Boughanem. DTD based costs for Tree-Edit distance in Structured Information Retrieval . In *European Conference on Information Retrieval (ECIR), Moscou, Russie*, pages 158–179. Springer, mars 2013a. https://doi.org/10.1007/978-3-642-36973-5_14. *6 citations pages iv, 12, 13, 16, 25, et IV*

Cyril Laitang, Karen Pinel-Sauvagnat, and Mohand Boughanem. Estimating Structural Relevance of XML Elements Through Language Model. In *Open Areas in Information Retrieval (OAIR), Lisbon, Portugal*. Kent State UNiversity, mai 2013b. http://oatao.univ-toulouse.fr/12412/. *7 citations pages iv, 12, 16, 17, 18, 25, et IV*

Mounia Lalmas. *XML Retrieval*. Synthesis Lectures on Information Concepts, Retrieval, and Services. Morgan & Claypool Publishers, 2009. http://dx.doi.org/10.2200/S00203ED1V01Y200907ICR007. *Cité page 8*

Mounia Lalmas. Aggregated search. In Massimo Melucci and Ricardo Baeza-Yates, editors, *Advanced Topics on Information Retrieval*. Springer, 2011. *2 citations pages 4 et 78*

Sylvain Lamprier. *Vers la conception de documents composites : Extraction et organisation de l'information pertinente*. PhD thesis, Université d'Angers, 2008. *Cité page 74*

Sylvain Lamprier, Tassadit Amghar, Bernard Levrat, and Frédéric Saubion. Organiser les résultats d'une recherche d'information – clustering, répartition de l'information et facilité d'accès. *Document NumÈrique*, 13(1) :9–39, 2010. *3 citations pages 72, 73, et 74*

Victor Lavrenko and W. Bruce Croft. 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. ACM, 2001. *Cité page 49*

Mai Dong Le. Graphes et appariement sémantique dedocuments XML. Rapport de master, IRIT, Université Paul Sabatier, Toulouse, Juin 2009. *Cité page 26*

Mai Dong Le and Karen Pinel-Sauvagnat. Utilisation de la distance d'édition pour l'appariement sémantique de documents XML . In *Atelier GAOC - Conférence EGC, Hammamet, Tunisie*. Association Internationale Francophone d'Extraction et de Gestion des Connaissances (EGC), janvier 2010. http://www.irit.fr/~Karen.Pinel-Sauvagnat/fichiers/GAOC2010.pdf. *Cité page 12*

Anton Leuski. Evaluating Document Clustering for Interactive Information Retrieval. In *CIKM'01*, pages 33–40. ACM, 2001. *2 citations pages 73 et 74*

VI Levenshtein. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, 10 :707, 1966. *Cité page 14*

David D. Lewis, Yiming Yang, Tony G. Rose, and Fan Li. Rcv1 : A new benchmark collection for text categorization research. *J. Mach. Learn. Res.*, 5 :361–397, 2004. *Cité page 75*

Fei-Fei Li, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples : an incremental bayesian approach tested on 101 object categories. In *IEEE Workshop on Generative-Model Based Vision*, 2004. *Cité page 77*

Rongmei Li and Theo P. van der Weide. Extended language models for xml element retrieval. In *Comparative Evaluation of Focused Retrieval - 9th International Workshop of the Inititative for the Evaluation of XML Retrieval, INEX 2010*, pages 89–97, 2010. *Cité page 16*

Xiao Li, Ye-Yi Wang, and Alex Acero. Learning query intent from regularized click graphs. In *SIGIR '08 : Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 339–346. ACM, 2008. ISBN 978-1-60558-164-4. doi : http://doi.acm.org/10.1145/1390334.1390393. *Cité page 78*

Xiaonan Li, Chengkai Li, and Cong Yu. Entity-relationship queries over wikipedia. *ACM Trans. Intell. Syst. Technol.*, 3(4) :70 :1–70 :20, September 2012. https://dl.acm.org/citation.cfm?id=2337555. *Cité page 42*

Girija Limaye, Sunita Sarawagi, and Soumen Chakrabarti. Annotating and searching web tables using entities, types and relationships. *Proc. VLDB Endow.*, 3 :1338–1347, September 2010. https://doi.org/10.14778/1920841.1921005. *Cité page 42*

Jimmy Lin and Miles Efron. Overview of the trec-2013 microblog track. In *Text REtrieval Conference (TREC), Gaithersburg, Maryland, USA*, 2013. *Cité page V*

Jimmy Lin, Miles Efron, Yulu Wang, Garrick Sherman, and Ellen Voorhees. Overview of the trec-2015 microblog track. In *Text REtrieval Conference (TREC), Gaithersburg, Maryland, USA*, 2015. *Cité page V*

Jimmy Lin, Adam Roegiest, Luchen Tan, Richard McCreadie, Ellen Voorhees, and Fernando Diaz. Overview of the trec 2016 real-time summarization track. In *Text REtrieval Conference (TREC), Gaithersburg, Maryland, USA*, 2016. *3 citations pages 81, 82, et V*

Jimmy Lin, Salman Mohammed, Royal Sequiera, Luchen Tan, Nimesh Ghelani, Mustafa Abualsaud, Richard McCreadie, Dmitrijs Milajevs, and Ellen Voorhees. Overview of the trec 2017 real-time summarization track. In *Text REtrieval Conference (TREC), Gaithersburg, Maryland, USA*, 2017. *3 citations pages 81, 82, et 84*

Thomas Lin and Oren Etzioni. Identifying functional relations in web text. In *Proc. of EMNLP 2010*, 2010. *Cité page 42*

Xiao Ling, Sameer Singh, and Daniel S Weld. Design challenges for entity linking. *Transactions of the Association for Computational Linguistics*, 3 :315–328, 2015. *Cité page 41*

Ning Liu, Jun Yan, and Zheng Chen. A probabilistic model based approach for blended search. In *WWW '09 : Proceedings of the 18th international conference on World wide web*, pages 1075–1076. ACM, 2009. *Cité page 78*

Qian Liu, Yue Liu, Dayong Wu, and Xueqi Cheng. Ictnet at temporal summarization track trec 2013. In *Text REtrieval Conference (TREC), Gaithersburg, Maryland, USA*, 2013a. *2 citations pages 59 et 60*

Tie-Yan Liu. *Learning to Rank for Information Retrieval*. Springer, 2011. *Cité page 94*

Xitong Liu, Jerry Darko, and Hui Fang. A related entity based approach for knowledge base acceleration. In *Text REtrieval Conference (TREC), Gaithersburg, Maryland, USA*, 2013b. *Cité page 43*

D. MacKay and L. Bauman Peto. A hierarchical dirichlet language model. *Natural Language Engineering*, 1 :1–19, 1994. *Cité page 18*

Matteo Magnani, Danilo Montesi, and Luca Rossi. Conversation retrieval for microblogging sites. *Inf. Retr.*, 15(3-4) :354–372, 2012. https://doi.org/10.1007/s10791-012-9189-9. *2 citations pages 32 et 34*

Saadia Malik, Mounia Lalmas, and Norbert Fuhr. *Overview of INEX 2004*. Springer, 2004. *Cité page 11*

Saadia Malik, Mounia Lalmas, and Norbert Fuhr. Overview of inex 2004. In Norbert Fuhr, Mounia Lalmas, Saadia Malik, and Zoltán Szlávik, editors, *Advances in XML Information Retrieval : Third International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2004, Dagstuhl Castle, Germany, December 6-8, 2004, Revised Selected Papers*, pages 1–15. Springer Berlin Heidelberg, 2005. *Cité page III*

Saadia Malik, Gabriella Kazai, Mounia Lalmas, and Norbert Fuhr. Overview of INEX 2005. In Norbert Fuhr, Mounia Lalmas, Saadia Malik, and Gabriella Kazai, editors, *Advances in XML Information Retrieval and Evaluation : 4th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2005, Dagstuhl Castle, Germany, November 28-30, 2005. Revised Selected Papers*, pages 1–15. Springer Berlin Heidelberg, 2006. *2 citations pages III et IV*

Saadia Malik, Andrew Trotman, Mounia Lalmas, and Norbert Fuhr. Overview of inex 2006. In Norbert Fuhr, Mounia Lalmas, and Andrew Trotman, editors, *Comparative Evaluation of XML Information Retrieval Systems : 5th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2006, Dagstuhl Castle, Germany, December 17-20, 2006, Revised and Selected Papers*, pages 1–11. Springer Berlin Heidelberg, 2007. *Cité page III*

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008. ISBN ISBN-13 978-0-521-86571-5. *Cité page 71*

Yossi Mass and Matan Mandelbrod. Retrieving the most relevant XML components. In *Proceedings of INEX 2003, Dagstuhl, Germany*, 2003. *Cité page 10*

Kamran Massoudi, Manos Tsagkias, Maarten de Rijke, and Wouter Weerkamp. Incorporating query expansion and quality indicators in searching microblog posts. In *Proceedings of the 33rd European conference on Advances in information retrieval*, ECIR'11, pages 362–367. Springer-Verlag, 2011. *Cité page 32*

Richard McCreadie, Craig Macdonald, and Iadh Ounis. Incremental update summarization : Adaptive sentence selection based on prevalence and novelty. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pages 301–310. ACM, 2014. *Cité page 60*

Xiaofeng Meng, Haiyan Wang, Dongdong Hu, and Chen Li. A supervised visual wrapper generator for web-data extraction. In *Proc. of COMPSAC '03*, page 657, 2003. *Cité page 42*

Donald Metzler and Congxing Cai. USC/ISI at TREC 2011 : Microblog Track. In *Text REtrieval Conference (TREC), Gaithersburg, Maryland, USA*. NIST, 2011. *3 citations pages 32, 33, et 34*

Diana Moise, Denis Shestakov, Gylfi Gudmundsson, and Laurent Amsaleg. Terabyte-scale image similarity search : Experience and best practice. In *Proceedings of the 2013 IEEE International Conference on Big Data, 6-9 October 2013, Santa Clara, CA, USA*, pages 674–682, 2013a. *Cité page 78*

Diana Moise, Denis Shestakov, Gylfi Gudmundsson, and Laurent Amsaleg. Indexing and searching 100m images with map-reduce. In *International Conference on Multimedia Retrieval, ICMR'13, Dallas, TX, USA, April 16-19, 2013*, pages 17–24, 2013b. *Cité page 78*

Véronique Moriceau and Xavier Tannier. FIDJI : using syntax for validating answers in multiple documents. *Information Retrieval*, 13 :507–533, October 2010. http://dx.doi.org/10.1007/s10791-010-9131-y. *2 citations pages 3 et 54*

Josiane Mothe, Jacques Savoy, Jaap Kamps, Karen Pinel-Sauvagnat, Gareth J.F. Jones, Eric San Juan, Linda Cappellato, and Nicola Ferro, editors. *Experimental IR Meets Multilinguality, Multimodality, and Interaction. 6th International Conference of the CLEF Association, CLEF'15, September 8-11, 2015, Proceedings, Toulouse, France, 08/09/2015 - 11/09/2015*, septembre 2015. Springer. http://www.springer.com/fr/book/9783319240268, http://doi.org/10.1007/978-3-319-24027-5. *Cité page 85*

Bilel Moulahi, Lamjed Ben Jabeur, Abdelhamid Chellal, Thomas Palmer, Lynda Tamine, Mohand Boughanem, Karen Pinel-Sauvagnat, and Gilles Hubert. IRIT at TREC Real Time Summarization 2016. In *Text REtrieval Conference (TREC), Gaithersburg, Maryland USA*, 2016. http://trec.nist.gov/pubs/trec25/papers/IRIT-RT.pdf. *4 citations pages 36, 37, 71, et V*

Pierre-Alain Moëllic and Christian Fluhr. Imageval 2006 official campaign, http://www.imageval.org. Voir aussi : http://cmm.ensmp.fr/~marcoteg/ImagEval_e.htm, 2006. *2 citations pages 76 et 77*

Henning Müller, Thomas Deselaers, Thomas Martin Deserno, Jayashree Kalpathy-Cramer, Eugene Kim, and William R Hersh. Overview of the ImageCLEFmed 2007 Medical Retrieval and Medical Annotation Tasks. In *CLEF*, pages 472–491. Springer, 2007. *Cité page IV*

Henning Müller, Jayashree Kalpathy-Cramer, Charles E Kahn Jr, William Hatt, Steven Bedrick, and William R Hersh. Overview of the imageclefmed 2008 medical image retrieval task. In *CLEF*, pages 512–522. Springer, 2008. *Cité page IV*

Vanessa Murdock and Mounia Lalmas. Workshop on aggregated search. *SIGIR Forum*, 42(2) :80–83, 2008. http://doi.acm.org/10.1145/1480506.1480520. *Cité page 2*

Henning Müller and Jayashree Kalpathy-Cramer. The medical image retrieval task. In *ImageClef, Experimental Evaluation in Visual Information Retrieval*, 2010. *Cité page 76*

Henning Müller, Wolfgang Müller, David McG. Squire, Stéphane Marchand-Maillet, and Thierry Pun. Performance evaluation in content-based image retrieval : overview and proposals. *Pattern Recognition Letters*, 22(5) :593 – 601, 2001. https://doi.org/10.1016/S0167-8655(00)00118-5. *Cité page 76*

Mor Naaman, Yee Jiun Song, Andreas Paepcke, and Hector Garcia-Molina. Assigning textual names to sets of geographic coordinates. *Computers, Environment and Urban Systems*, 30(4) :418–435, 2006. https://doi.org/10.1016/j.compenvurbsys.2006.02.001. *Cité page 55*

Rinkesh Nagmoti, Ankur Teredesai, and Martine De Cock. Ranking approaches for microblog search. In *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on*, volume 1, pages 153–157. IEEE, 2010. *2 citations pages 33 et 37*

Emmanuel Navarro. *Metrology of terrain networks, application to the construction of lexical resources and to information retrieval.* PhD thesis, Institut National Polytechnique de Toulouse - INPT, November 2013. URL https://tel.archives-ouvertes.fr/tel-01020232. *Cité page 75*

Emmanuel Navarro, Yannick Chudy, Bruno Gaume, Guillaume Cabanac, and Karen Pinel-Sauvagnat. Kodex ou comment organiser les résultats d'une recherche d'information par détection de communautés sur un graphe biparti ? . In *Conférence francophone en Recherche d'Information et Applications (CORIA), Avignon*, pages 25–40. Association Francophone de Recherche d'Information et Applications (ARIA), mars 2011. http://doi.org/doi:10.24348/coria.2011.25. *Cité page 75*

Richi Nayak, Christopher M. De Vries, Sangeetha Kutty, Shlomo Geva, Ludovic Denoyer, and Patrick Gallinari. Overview of the inex 2009 XML mining track : Clustering and classification of xml documents. In *INEX*, pages 366–378, 2009. *Cité page 73*

Ani Nenkova, Kathleen McKeown, et al. Automatic summarization. *Foundations and Trends in Information Retrieval*, 5(2–3) :103–233, 2011. http://dx.doi.org/10.1561/1500000015. *Cité page 59*

Zaiqing Nie, Y unxiaoMa, Shuming Shi, Ji-Rong Wen, and Wey-Ying Ma. Web object retrieval. In *WWW 7*, pages 81–90, 2007. *Cité page 54*

Shahrul Noah, A. Azilawati, Tengku Sembok, and Siti Meriam. Exploiting surrounding text for retrieving web images. *Journal of Computer Science*, pages 842–846, 2008. http://doi.org/10.3844/jcssp.2008.842.846. *Cité page 21*

Stefanie Nowak, Allan Hanbury, and Thomas Deselaers. Object and concept recognition for image retrieval. In *ImageClef, Experimental Evaluation in Visual Information Retrieval*, 2010. *Cité page 77*

Paul Ogilvie and Jamie Callan. Hierarchical language models for xml component retrieval. In Norbert Fuhr, Mounia Lalmas, Saadia Malik, and Zoltán Szlávik, editors, *Advances in XML Information Retrieval*, volume 3493 of *Lecture Notes in Computer Science*, pages 269–285. Springer Berlin / Heidelberg, 2005. *Cité page 16*

Iadh Ounis, Gianni Amati, Vassilis Plachouras, Ben He, Craig Macdonald, and Douglas Johnson. Terrier information retrieval platform. In *European Conference on Information Retrieval*, pages 517–519. Springer, 2005. *Cité page 75*

Iadh Ounis, Jimm Lin, and Ian Soboroff. Overview of the TREC-2011 Microblog Track. In *Text REtrieval Conference (TREC), Gaithersburg, Maryland, USA*, 2011. *2 citations pages 29 et V*

Iadh Ounis, Jimm Lin, and Ian Soboroff. Overview of the TREC-2012 Microblog Track. In *Text REtrieval Conference (TREC), Gaithersburg, Maryland, USA*, 2012. *Cité page V*

Alexander Pak and Patrick Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. In *LREc*, 2010. *Cité page 29*

Damien Palacio, Guillaume Cabanac, Gilles Hubert, Karen Pinel-Sauvagnat, and Christian Sallaberry. Prototyping a Personalized Contextual Retrieval Framework. In *ACM SIGSPATIAL Workshop on Geographic Information Retrieval (GIR), Orlando, USA*, pages 43–44. ACM, novembre 2013. http://doi.org/10.1145/2533888.2533935. *Cité page 91*

Thomas Palmer, Gilles Hubert, and Karen Pinel-Sauvagnat. Retweeter ou ne pas retweeter : le dilemme des portails de diffusion d'information temps-réel . In *Conférence francophone en Recherche d'Information et Applications (CORIA), Marseille*, pages 123–138. Association Francophone de Recherche d'Information et Applications (ARIA), mars 2017. http://doi.org/doi:10.24348/coria.2017.28. *4 citations pages 36, 37, 38, et V*

Monica Lestari Paramita and Michael Grubinger. Photographic image retrieval. In *ImageClef, Experimental Evaluation in Visual Information Retrieval*, 2010. *Cité page 76*

Cécile Paris, Stephen Wan, and Paul Thomas. Focused and aggregated search : A perspective from natural language generation. *Inf. Retr.*, 13(5) :434–459, 2010. http://dx.doi.org/10.1007/s10791-009-9121-0. *2 citations pages 3 et 54*

Marius Paşca and Benjamin Van Durme. Weakly-supervised acquisition of open-domain classes and class attributes from web documents and query logs. In *Proceedings of ACL-08 : HLT*, pages 19–27, 2008. *Cité page 46*

Fabien Petitcolas, Ross J. Anderson, and Markus G. Kuhn. Attacks on copyright marking systems. In *Second International Workshop IH-98, Portland, Oregon, USA*, pages 219–239, 1998. *Cité page 77*

Karen Pinel-Sauvagnat. Propagation-based structured text retrieval. In Ling Liu and M. Tamer Özsu, editors, *Encyclopedia of Database Systems*, pages 2197–2201. Springer, mai 2009. http://dx.doi.org/10.1007/978-0-387-39940-9_281. Sur invitation. *2 citations pages 10 et 11*

Karen Pinel-Sauvagnat. Propagation-Based Structured Text Retrieval. In Ling Liu and M. Tamer Özsu, editors, *Encyclopedia of Database Systems, revised edition*. Springer, août 2017. https://doi.org/10.1007/978-1-4899-7993-3_281-2. *Cité page 10*

Karen Pinel-Sauvagnat and Mohand Boughanem. Propositions pour la pondération des termes et l'évaluation de la pertinence des éléments en recherche d'information structurée. *Information - Interaction - Intelligence*, 6(2) :77–98, décembre 2006. http://www.irit.fr/~Karen.Pinel-Sauvagnat/fichiers/I32006.pdf. *2 citations pages 12 et III*

Karen Pinel-Sauvagnat and Mohand Boughanem. A survey on XML focussed component retrieval . In *Large-Scale Semantic Access to Content (Text, Image, Video and Sound) (RIAO), Pittsburgh*. Centre de hautes études internationales d'Informatique Documentaire (C.I.D.), juin 2007. http://www.irit.fr/~Karen.Pinel-Sauvagnat/fichiers/RIAO2007.pdf. *2 citations pages 12 et III*

Karen Pinel-Sauvagnat and Claude Chrisment. XML et recherche d'information. In Mohand Boughanem and Jacques Savoy, editors, *Recherche d'information. Etat des lieux et perspectives.*, chapter 4, pages 99–138. Hermès, http://www.editions-hermes.fr/, avril 2008. https://www.lavoisier.fr/livre/informatique/recherche-d-information-etat-des-lieux-et-perspectives/boughanem/descriptif-9782746220058, http://www.irit.fr/~Karen.Pinel-Sauvagnat/fichiers/Hermes2008.pdf, Sur invitation. *Cité page 8*

Benjamin Piwowarski, Georges-Etienne Faure, and Patrick Gallinari. Bayesian networks and INEX. In *Proceedings in the First INEX Workshop*, December 2002. *Cité page 10*

Jay M. Ponte and W. Bruce Croft. 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. ACM, 1998. *Cité page 48*

Ana-Maria Popescu and Oren Etzioni. Extracting product features and opinions from reviews. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 339–346. Association for Computational Linguistics, 2005. doi : 10.3115/1220575.1220618. URL https://doi.org/10.3115/1220575.1220618. *Cité page 45*

Eugen Popovici, Gildas Menier, and Pierre-Francois Marteau. SIRIUS : A lightweight XML indexing and approximate search system at INEX 2005. In *Proceedings of the Initiative for the Evaluation of XML Retrieval*, pages 321–335, 2005. *Cité page 14*

Jeffrey Pound, Peter Mika, and Hugo Zaragoza. 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. ACM, 2010. *Cité page 40*

Xin Qian, Jimmy Lin, and Adam Roegiest. Interleaved evaluation for retrospective summarization and prospective notification on document streams. In *Proceedings of the 39th International ACM SIGIR Conference*, SIGIR '16, pages 175–184, 2016. *Cité page 84*

Roy Rada, Hafedh Mili, Ellen Bicknell, and Maria Blettner. Development and application of a metric on semantic nets. *IEEE Transactions on Systems, Man, and Cybernetics*, 19(1) :17–30, 1989. 10.1109/21.24528. *Cité page 24*

Stephen E. Robertson. *The probability ranking principle in IR*, pages 281–286. Readings in information retrieval, Morgan Kaufmann Publishers Inc., 1997. ISBN 1-55860-454-5. URL http://portal.acm.org/citation.cfm?id=275537.275701. *Cité page 62*

Stephen E Robertson and K Sparck Jones. Relevance weighting of search terms. *Journal of the American Society for Information science*, 27(3) :129–146, 1976. https://doi.org/10.1002/asi.4630270302. *Cité page 18*

Joseph John Rocchio. Relevance feedback in information retrieval. *The SMART retrieval system : experiments in automatic document processing*, pages 313–323, 1971. *2 citations pages 18 et 31*

Adam Roegiest, Luchen Tan, and Jimmy Lin. Online in-situ interleaved evaluation of real-time push notification systems. In *Proceedings of the 40th International ACM SIGIR Conference*, SIGIR '17, pages 415–424, 2017. *Cité page 81*

Cyril Rohr and Dian Tjondronegoro. Aggregated cross-media news visualization and personalization. In *Proc. of MIR 2008, Vancouver, British Columbia, Canada*, pages 371–378, 2008. *Cité page 55*

Nachiketa Sahoo, Jamie Callan, Ramayya Krishnan, George Duncan, and Rema Padman. Incremental hierarchical clustering of text documents. In *Proc. of CIKM 2006, Arlington, Virginia, USA*, pages 357–366, 2006. *Cité page 55*

Taro L. Saito and Shinichi Morishita. Amoeba Join : Overcoming Structural Fluctuations in XML Data. In *WebDB*, 2006. http://db.ucsd.edu/webdb2006/camera-ready/paginated/08-110.pdf. *Cité page 14*

Gerard Salton and Chris Buckley. Improving retrieval performance by relevance feedback. In Karen Sparck Jones and Peter Willett, editors, *Readings in information retrieval*, pages 355–364. Morgan Kaufmann Publishers Inc., 1997. ISBN 1-55860-454-5. *2 citations pages 18 et 31*

Mark Sanderson. Test collection based evaluation of information retrieval systems. *Foundations and Trends in Information Retrieval*, 4(4) :247–375, 2010a. http://dx.doi.org/10.1561/1500000009. *2 citations pages 71 et 95*

Mark Sanderson. Performance measures used in image information retrieval. In *ImageClef, Experimental Evaluation in Visual Information Retrieval*, 2010b. *Cité page 76*

Tefko Saracevic. Evaluation of evaluation in information retrieval. In *Proc. ACM SIGIR conference on Research and development in information retrieval*, pages 138–146, 1995. *Cité page 71*

Karen Sauvagnat. *Modèle flexible pour la recherche d'information dans des corpus de documents semi-structurés.* Thèse de doctorat, Université Paul Sabatier, Toulouse, France, juin 2005. https://hal-univ-tlse3.archives-ouvertes.fr/tel-00359579, (Soutenance le 30/06/2005). *2 citations pages 5 et 11*

Karen Sauvagnat and Mohand Boughanem. Propositions pour la pondération des termes et l"évaluation de la pertinence des éléments en recherche d'information structurée . In *Conférence francophone en Recherche d'Information et Applications (CORIA), Lyon*, pages 29–40. Association Francophone de Recherche d'Information et Applications (ARIA), mars 2006. http://doi.org/doi:10.24348/coria.2006.29. *2 citations pages 12 et III*

Karen Sauvagnat, Lobna Hlaoua, and Mohand Boughanem. XFIRM at INEX 2005 : adhoc and relevance feedback tracks. In Norbert Fuhr, Mounia Lalmas, Saadia Malik, and Gabriella Kazai, editors, *INitiative for the Evaluation of XML Retrieval (INEX), Dagstuhl, Germany*, volume 3977 of *LNCS*, pages 88–103. Springer, novembre 2005. https://doi.org/10.1007/978-3-540-34963-1_7. *4 citations pages 12, 18, 25, et III*

Karen Sauvagnat, Mohand Boughanem, and Claude Chrisment. Why using structural hints in XML retrieval ? . In Henrik Legind Larsen, Gabriella Pasi, and Ortiz-Arroyo Daniel, editors, *Flexible Query Answering (FQAS), Milan, Italie*, Advances in Artificial Intelligence, pages 197–109. Springer, juin 2006a. https://doi.org/10.1007/11766254_17. *2 citations pages 12 et III*

Karen Sauvagnat, Mohand Boughanem, and Claude Chrisment. Answering content-and-structure-based queries on XML documents using relevance propagation. *Information Systems, Special Issue SPIRE 2004*, 31 :621–635, janvier 2006b. http://dx.doi.org/10.1016/j.is.2005.11.007. *2 citations pages 12 et III*

Karen Sauvagnat, Lobna Hlaoua, and Mohand Boughanem. XML retrieval : what about using contextual relevance ? In *Annual ACM Symposium on Applied Computing (SAC), Dijon*, pages 1114–1120. ACM Press, avril 2006c. https://doi.org/10.1007/11766254_17. *2 citations pages 12 et III*

Torsten Schlieder and Holgers Meuss. Querying and ranking XML documents. *Journal of the American Society for Information Science and Technology*, 53(6) :pages 489–503, 2002. http://doi.org/10.1002/asi.10060. *2 citations pages 10 et 12*

Erik Selberg and Oren Etzioni. Multi-service search and comparison using the metacrawler. In *Proceedings of the Fourth Int'l WWW Conference, Boston*, 1995. *Cité page 54*

W. Shen, J. Wang, and J. Han. Entity linking with a knowledge base : Issues, techniques, and solutions. *IEEE Transactions on Knowledge and Data Engineering*, 27(2) :443–460, Feb 2015. http ://doi.org/10.1109/TKDE.2014.2327028. *Cité page 41*

Amit Singhal. Introducing the knowledge graph : things, not strings, may 2012. https://search.googleblog.com/2012/05/introducing-knowledge-graph-things-not.html. *2 citations pages 3 et 57*

Ian Soboroff and Donna Harman. Novelty detection : the trec experience. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 105–112. Association for Computational Linguistics, 2005. *Cité page 59*

Herbert A. Sturges. The Choice of a Class Interval. *Journal of the American Statistical Association*, 21(153) :65–66, 1926. *Cité page 34*

Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. YAGO : A large ontology from Wikipedia and Wordnet. *Web Semant.*, 6(3) :203–217, 2008. http://dx.doi.org/10.1016/j.websem.2008.06.001. *Cité page 42*

Fabian M. Suchanek, Mauro Sozio, and Gerhard Weikum. Sofie : a self-organizing framework for information extraction. In *Proceedings of WWW Conference*, pages 631–640, 2009. *Cité page 41*

Mihai Surdeanu and Heng Ji. Overview of the english slot filling track at the tac2014 knowledge base population evaluation. In *Proc. Text Analysis Conference (TAC2014)*, 2014. *2 citations pages 42 et 59*

Shanu Sushmita, Hideo Joho, and Mounia Lalmas. A task-based evaluation of an aggregated search interface. In *SPIRE '09 : Proceedings of the 16th International Symposium on String Processing and Information Retrieval*, pages 322–333. Springer-Verlag, 2009. *Cité page 78*

Shanu Sushmita, Hideo Joho, Mounia Lalmas, and Robert Villa. Factors affecting click-through behavior in aggregated search interfaces. In *CIKM '10 : Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 519–528. ACM, 2010. ISBN 978-1-4503-0099-5. doi : http://doi.acm.org/10.1145/1871437.1871506. *Cité page 78*

Yu Suzuki and Masatoshi Yoshikawa. Assessing quality score of wikipedia article using mutual evaluation of editors and texts. In *Proceedings of the 22Nd ACM International Conference on Conference on Information and Knowledge Management*, CIKM '13, pages 1727–1732, 2013. *Cité page 42*

Mohammed Amin Tahraoui, Karen Pinel-Sauvagnat, Cyril Laitang, Mohand Boughanem, Hamamache Kheddouci, and Lei Ning. A survey on tree matching and XML retrieval. *Computer Science Review*, 8 :1–23, mai 2013. http://dx.doi.org/10.1016/j.cosrev.2013.02.001. *Cité page 12*

Kuo-Chung Tai. The tree-to-tree correction problem. *J. ACM*, 26 :422–433, July 1979. https://doi.org/10.1145/322139.322143. *Cité page 14*

Lynda Tamine-Lechani, Mohand Boughanem, and Mariam Daoud. Evaluation of contextual information retrieval effectiveness : overview of issues and research. *Knowledge and Information Systems*, 24(1) :1–34, 2009. http://dx.doi.org/10.1007/s10115-009-0231-1. *Cité page 2*

Luchen Tan, Adam Roegiest, Jimmy Lin, and Charles L.A. Clarke. An exploration of evaluation metrics for mobile push notifications. In *Proceedings of the 39th International ACM SIGIR Conference*, SIGIR '16, pages 741–744, 2016. *2 citations pages 81 et 82*

Luchen Tan, Gaurav Baruah, and Jimmy Lin. On the reusability of "living labs" test collections : A case study of real-time summarization. In *Proceedings of the 40th International ACM SIGIR Conference*, SIGIR '17, pages 793–796, 2017. *Cité page 81*

Jaime Teevan, Daniel Ramage, and Merredith Ringel Morris. #twittersearch : a comparison of microblog search and web search. In *WSDM'11 : Proceedings of the fourth ACM international conference on Web search and data mining*, pages 35–44. ACM, 2011. ISBN 978-1-4503-0493-1. *2 citations pages 29 et 36*

Paul Thomas, Katherine Noack, and Cecile Paris. Evaluating interfaces for government metasearch. In *Proceedings of the third symposium on Information interaction in context*, pages 65–74. ACM, 2010. *Cité page 78*

Thibaut Thonet. *Modèles thématiques pour la découverte non supervisée de points de vue sur le Web*. Thèse de doctorat, Université Paul Sabatier, Toulouse, France, décembre 2017. https://tel.archives-ouvertes.fr/tel-01655278, Soutenance le 23/11/2017. *Cité page 90*

Thibaut Thonet, Guillaume Cabanac, Mohand Boughanem, and Karen Pinel-Sauvagnat. VODUM : a Topic Model Unifying Viewpoint, Topic and Opinion Discovery . In *European Conference on Information Retrieval (ECIR), Padua, Italy*, volume 9626 of *LNCS*, pages 533–545. Springer, mars 2016. http://doi.org/10.1007/978-3-319-30671-1_39. *2 citations pages 90 et 91*

Thibaut Thonet, Guillaume Cabanac, Mohand Boughanem, and Karen Pinel-Sauvagnat. Users Are Known by the Company They Keep : Topic Models for Viewpoint Discovery in Social Networks . In *Conference on Information and Knowledge Management (CIKM), Singapore*, pages 87–96. ACM, novembre 2017. https://doi.org/10.1145/3132847.3132897. *2 citations pages 90 et 91*

Kosuke Tokunaga, Jun'ichi Kazama, and Kentaro Torisawa. Automatic discovery of attribute words from web documents. In Robert Dale, Kam-Fai Wong, Jian Su, and Oi Yee Kwong, editors, *Natural Language Processing – IJCNLP 2005*, pages 106–118. Springer Berlin Heidelberg, 2005. *Cité page 46*

Anastasios Tombros. *The effectiveness of hierarchic query-based clustering of documents for information retrieval*. Thèse de doctorat, University of Glasgow, UK, 2002. *2 citations pages 72 et 73*

Mouna Torjmen. Recherche contextuelle d'images dans des documents XML. Rapport de master, IRIT, Université Paul Sabatier, Toulouse, juillet 2006. *Cité page 26*

Mouna Torjmen. *Approches de Recherche Multimedia dans des Documents Semi-Structurés : Utilisation du contexte textuel et structurel pour la sélection d'objets multimedia*. Thèse de doctorat, Université Paul Sabatier, Toulouse, France, décembre 2009. http://oatao.univ-toulouse.fr/12616/1/Torjmen-Khemakhem_12616.pdf, Soutenance le 04/12/2009. *2 citations pages 26 et 89*

Mouna Torjmen and Karen Pinel-Sauvagnat. Une étude de l'impact de la structure sur la recherche multimedia . In *Conférence francophone en Recherche d'Information et Applications (CORIA), Presqu'île de Giens- Var*, pages 51–66. Ludovia, mai 2009. http://doi.org/doi:10.24348/coria.2009.51. *3 citations pages 23, 24, et IV*

Mouna Torjmen, Karen Pinel-Sauvagnat, and Mohand Boughanem. MM-XFIRM at INEX Multimedia track 2007 (working notes), décembre 2007a. http://inex.mmci.uni-saarland.de/static/proceedings/INEX2007-preproceedings.pdf. *3 citations pages 25, 71, et IV*

Mouna Torjmen, Karen Pinel-Sauvagnat, and Mohand Boughanem. Utilisation du contexte textuel et structurel pour la recherche d'images dans des documents XML . In *Congrès Informatique des Organisations et Systèmes d'Information et de Décision (INFORSID), Perros-Guirec*, pages 21–36. IRISA, mai 2007b. *2 citations pages 23 et IV*

Mouna Torjmen, Karen Pinel-Sauvagnat, and Mohand Boughanem. Towards a structure-based multimedia retrieval model. In *ACM International Conference on Multimedia Information Retrieval , Vancouver, Canada*, pages 350–357. ACM, octobre 2008a. http://dx.doi.org/10.1145/1460096.1460153. *3 citations pages 23, 24, et IV*

Mouna Torjmen, Karen Pinel-Sauvagnat, and Mohand Boughanem. Une métrique pondérée pour la recherche textuelle d'images dans des documents semi-structurés . In *Conférence francophone en Recherche d'Information et Applications (CORIA), Trégastel*, pages 55–70. Université de Rennes 1, mars 2008b. http://doi.org/doi:10.24348/coria.2008.55. *3 citations pages 23, 24, et IV*

Mouna Torjmen, Karen Pinel-Sauvagnat, and Mohand Boughanem. Methods for combining content-based and textual-based approaches in medical image retrieval(working notes), septembre 2008c. Preproceeding de CLEF 2008, pas de sélection. *Cité page 71*

Mouna Torjmen, Karen Pinel-Sauvagnat, and Mohand Boughanem. Some experiments on the WikipediaMM 2008 task : Evaluating the impact of image names in context-based retrieval (working notes), septembre 2008d. Preproceeding de CLEF 2008, pas de sélection. *Cité page 71*

Mouna Torjmen, Karen Pinel-Sauvagnat, and Mohand Boughanem. Using pseudo-relevance feedback to improve image retrieval results. In Alessandro Nardi and Carol Peters, editors, *Advances in Multilingual and Multimodal Information Retrieval. CLEF 2007, Budapest, Hungary*, LNCS, pages 665–673. Springer, septembre 2008e. https://doi.org/10.1007/978-3-540-85760-0_85. *4 citations pages 21, 25, 71, et IV*

Mouna Torjmen, Karen Pinel-Sauvagnat, and Mohand Boughanem. Methods for combining content-based and textual-based approaches in medical image retrieval. In *Evaluating Systems for Multilingual and Multimodal Information Access. 9th Workshop of the Cross-Language Evaluation Forum, CLEF 2008, Aarhus, Denmark*, LNCS, pages 691–695. Springer, septembre 2009a. https://doi.org/10.1007/978-3-642-04447-2_87. *4 citations pages 21, 25, 71, et IV*

Mouna Torjmen, Karen Pinel-Sauvagnat, and Mohand Boughanem. Evaluating the impact of image names in context-based retrieval. In *Evaluating Systems for Multilingual and Multimodal Information Access. 9th Workshop of the Cross-Language Evaluation Forum, CLEF 2008, Aarhus, Denmark*, LNCS, pages 756–762. Springer, septembre 2009b. https://doi.org/10.1007/978-3-642-04447-2_98. *3 citations pages 25, 71, et IV*

Mouna Torjmen, Karen Pinel-Sauvagnat, and Mohand Boughanem. XML Multimedia Retrieval : From relevant textual information to relevant multimedia fragments . In *European Conference on Information Retrieval (ECIR), Toulouse*, pages 150–161. Springer, 2009c. http://dx.doi.org/10.1007/978-3-642-00958-7_16. *4 citations pages 23, 24, 25, et IV*

Mouna Torjmen, Karen Pinel-Sauvagnat, and Mohand Boughanem. Using textual and structural context for searching multimedia elements. *International Journal of Business Intelligence and Data Mining, Special Issue on Beyond Multimedia and XML Streams Querying and Mining*, 5(4) :323–352, octobre 2010. http://dx.doi.org/10.1504/IJBIDM.2010.036123. *3 citations pages 23, 24, et IV*

Mouna Torjmen-Khemakhem, Karen Pinel-Sauvagnat, and Mohand Boughanem. Investigating the document structure as a source of evidence for multimedia fragment retrieval. *Information Processing & Management*, 49 :1281–1300, juillet 2013. http://dx.doi.org/10.1016/j.ipm.2013.06.001. *3 citations pages 24, 25, et IV*

Thanh Tran, Peter Mika, Haofen Wang, and Marko Grobelnik. Semsearch'11 : the 4th semantic search workshop. In *Proceedings of the 20th international conference companion on World wide web*, pages 315–316. ACM, 2011. *Cité page 97*

Andrew Trotman. Processing structural constraints. In Ling Liu and M. Tamer Özsu, editors, *Encyclopedia of Database Systems*, pages 2191–2195. Springer, 2009. *Cité page 12*

Andrew Trotman and Mounia Lalmas. Why structural hints in queries do not help XML-retrieval. In *Proceedings of SIGIR*, pages 711–712, 2006. *Cité page 12*

Andrew Trotman and Börkur Sigurbjörnsson. NEXI, now and next. In *INEX 2003 proceedings, Dagstuhl, Allemagne*, pages 10–15, December 2004. *Cité page 10*

Andrew Trotman and Qiuyue Wang. Overview of the inex 2010 data centric track. In *Proceedings of the 9th International Conference on Initiative for the Evaluation of XML Retrieval : Comparative Evaluation of Focused Retrieval*, INEX'10, pages 171–181. Springer-Verlag, 2011. *Cité page IV*

Theodora Tsikrika. Aggregation-based structured text retrieval. In Ling Liu and M. Tamer Özsu, editors, *Encyclopedia of Database Systems*, pages 63–71. Springer, 2009. *Cité page 10*

Theodora Tsikrika and Jana Kludas. Overview of the wikipediamm task at imageclef 2008. In *In Evaluating Systems for Multilingual and Multimodal Information Access, Proceedings of the 9th Workshop of the Cross-Language Evaluation Forum, Lecture Notes in Computer Science*. Citeseer, 2009. *Cité page IV*

Theodora Tsikrika and Jana Kludas. The wikipedia image retrieval task. In *ImageClef, Experimental Evaluation in Visual Information Retrieval*, 2010. *Cité page 76*

Theodora Tsikrika and Thijs Westerveld. The inex 2007 multimedia track. In Norbert Fuhr, Jaap Kamps, Mounia Lalmas, and Andrew Trotman, editors, *Focused Access to XML Documents : 6th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2007 Dagstuhl Castle, Germany, December 17-19, 2007. Selected Papers*, pages 440–453. Springer Berlin Heidelberg, 2008. *2 citations pages 21 et IV*

Theodora Tsikrika, Pavel Serdyukov, Henning Rode, Thijs Westerveld, Robin Aly, Djoerd Hiemstra, and Arjen Vries. Structured Document Retrieval, Multimedia Retrieval, and Entity Ranking Using PF/Tijah. In *INEX*, pages 273–286, 2007. *Cité page 23*

Lauren Turpin, Diane Kelly, and Jaime Arguello. To blend or not to blend ? : Perceptual speed, visual memory and aggregated search. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 1021–1024. ACM, 2016. *Cité page 78*

David Vallet and Hugo Zaragoza. Inferring the most important types of a query : A semantic approach. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '08, pages 857–858, 2008. http://doi.acm.org/10.1145/1390334.1390541. *2 citations pages 4 et 55*

Roelof van Zwol, Gabriella Kazai, and Mounia Lalmas. Inex 2005 multimedia track. In Norbert Fuhr, Mounia Lalmas, Saadia Malik, and Gabriella Kazai, editors, *Advances in XML Information Retrieval and Evaluation : 4th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2005, Dagstuhl Castle, Germany, November 28-30, 2005. Revised Selected Papers*, pages 497–510. Springer Berlin Heidelberg, 2006. *Cité page IV*

Anne-Marie Vercoustre, Jovan Pehcevski, and James A. Thom. Using wikipedia categories and links in entity ranking. In *Proceedings of INEX*, pages 321–335, 2007. *Cité page 62*

Ellen Voorhees. The TREC-8 Question Answering track report. In *Text REtrieval Conference (TREC), Gaithersburg, Maryland, USA*, pages 77–82, 1999. *Cité page 77*

Jan Vosecky, Kenneth Wai-Ting Leung, and Wilfred Ng. Searching for quality microblog posts : Filtering and ranking based on content analysis and implicit links. In *DASFAA (1)*, pages 397–413, 2012. *Cité page 32*

Christopher Michael De Vries, Richi Nayak, Sangeetha Kutty, Shlomo Geva, and Andrea Tagarelli. Overview of the INEX 2010 XML mining track : clustering and classification of XML documents. In *Initiative for the Evaluation of XML Retrieval (INEX) 2010*. Springer, 2011. *Cité page 73*

W3C. EXtensible Markup Language (XML) 1.0. Technical report, World Wide Web Consortium (W3C), Recommandation, february 1998a. http://www.w3.org/TR/1998/REC-xml-19980210. *Cité page 8*

W3C. DOM Level 1 (Document Object Model). Technical report, World Wide Web Consortium (W3C), W3C Recommandation, october 1998b. http://www.w3c.org/DOM. *Cité page 8*

Jingang Wang, Dandan Song, Chin-Yew Lin, and Lejian Liao. Bit and msra at trec kba ccr track 2013. In *Text REtrieval Conference (TREC), Gaithersburg, Maryland, USA*, 2013. *Cité page 43*

Jingang Wang, Dandan Song, Qifan Wang, Zhiwei Zhang, Luo Si, Lejian Liao, and Chin-Yew Lin. 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*, pages 635–644. ACM, 2015a. *Cité page 43*

Qiuyue Wang, Georgina Ramírez, Maarten Marx, Martin Theobald, and Jaap Kamps. Overview of the inex 2011 data-centric track. In Shlomo Geva, Jaap Kamps, and Ralf Schenkel, editors, *Focused Retrieval of Content and Structure : 10th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2011, Saarbrücken, Germany, December 12-14, 2011, Revised Selected Papers*, pages 118–137. Springer Berlin Heidelberg, 2012. *Cité page IV*

Yulu Wang, Garrick Sherman, Jimmy Lin, and Miles Efron. Assessor differences and user preferences in tweet timeline generation. In *Proceedings of the 38th International ACM SIGIR Conference*, SIGIR '15, pages 615–624, 2015b. *Cité page 81*

Jianshu Weng, Ee-Peng Lim, Jing Jiang, and Qi He. Twitterrank : finding topic-sensitive influential twitterers. In *WSDM'10 : Proceedings of the third ACM international conference on Web search and data mining*, pages 261–270. ACM, 2010. *Cité page 29*

Thijs Westerveld and Roelof van Zwol. The inex 2006 multimedia track. In Norbert Fuhr, Mounia Lalmas, and Andrew Trotman, editors, *Comparative Evaluation of XML Information Retrieval Systems : 5th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2006, Dagstuhl Castle, Germany, December 17-20, 2006, Revised and Selected Papers*, pages 331–344. Springer Berlin Heidelberg, 2007a. *2 citations pages 21 et IV*

Thijs Westerveld and Roelof van Zwol. Multimedia retrieval at INEX 2006. *ACM SIGIR Forum*, 41 (1) :58–63, 2007b. https://doi.org/10.1145/1273221.1273226. *Cité page 11*

Fei Wu, Raphael Hoffmann, and Daniel S. Weld. Information extraction from wikipedia : moving down the long tail. In *Proc. of KDD 2008, Las Vegas, Nevada, USA*, pages 731–739, 2008. *Cité page 42*

Zhibiao Wu and Martha Palmer. Verbs semantics and lexical selection. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, ACL 94, pages 133–138, 1994. *Cité page 24*

Tan Xu, Douglas W Oard, and Paul McNamee. Hltcoe at trec 2013 : Temporal summarization. In *Text REtrieval Conference (TREC), Gaithersburg, Maryland, USA*, 2013. *2 citations pages 59 et 60*

Yuto Yamaguchi, Tsubasa Takahashi, Toshiyuki Amagasa, and Hiroyuki Kitagawa. Turank : Twitter user ranking based on user-tweet graph analysis. In Lei Chen, Peter Triantafillou, and Torsten Suel, editors, *Web Information Systems Engineering – WISE 2010*, pages 240–253. Springer Berlin Heidelberg, 2010. *Cité page 29*

Grace Hui Yang and Ian Soboroff. Trec 2016 dynamic domain track overview. In *Text REtrieval Conference (TREC), Gaithersburg, Maryland, USA*, 2016. *Cité page 95*

Show-Jane Yen and Yue-Shi Lee. Under-sampling approaches for improving prediction of the minority class in an imbalanced dataset. In *Intelligent Control and Automation*, volume 344 of *Lecture Notes in Control and Information Sciences*, pages 731–740. Springer Berlin / Heidelberg, 2006. *Cité page 34*

Naoki Yoshinaga and Kentaro Torisawa. Open-domain attribute-value acquisition from semi-structured texts. In *Proceedings of the 6th International Semantic Web Conference (ISWC-07), Workshop on Text to Knowledge : The Lexicon/Ontology Interface (OntoLex-2007)*, pages 55–66, 2007. *2 citations pages 45 et 46*

Haïfa Zargayouna. Contexte et sémantique pour une indexation de documents semi-structurés. In *Conférence francophone en Recherche d'Information et Applications (CORIA), Toulouse*, pages 571–581, 2004. *Cité page 24*

Chengxiang Zhai and John Lafferty. A study of smoothing methods for language models applied to information retrieval. *ACM Trans. Inf. Syst.*, 22(2) :179–214, 2004. https://doi.org/10.1145/984321.984322. *Cité page 18*

Chunyun Zhang, Weiyan Xu, Fanyu Meng, Hongyan Li, Tong Wu, and Lixin Xu. The information extraction systems of pris at temporal summarization track. In *Text REtrieval Conference (TREC), Gaithersburg, Maryland, USA*, 2013. *Cité page 60*

Le Zhao and Jamie Callan. Effective and efficient structured retrieval. In *CIKM*, pages 1573–1576, 2009. *Cité page 16*

Lulin Zhao, Yi Zeng, and Ning Zhong. A weighted multi-factor algorithm for microblog search. In *Proceedings of the 7th international conference on Active media technology*, AMT'11, pages 153–161. Springer-Verlag, 2011. *3 citations pages 32, 33, et 37*

Xiaoqi Zhao and Keishi Tajima. Online retweet recommendation with item count limits. In *Proceedings of the 2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)-Volume 01*, pages 282–289. IEEE Computer Society, 2014. *Cité page 37*

Yun Zhao, Fei Yao, Huayang Sun, and Zhen Yang. Bjut at trec 2014 temporal summarization track. In *Text REtrieval Conference (TREC), Gaithersburg, Maryland, USA*, 2014. *Cité page 59*

Yuxiang Zhao and Qinghua Zhu. Evaluation on crowdsourcing research : Current status and future direction. *Information Systems Frontiers*, 16(3) :417–434, 2014. https://doi.org/10.1007/s10796-012-9350-4. *Cité page 97*

Mianwei Zhou and Kevin Chen-Chuan Chang. Entity-centric document filtering : boosting feature mapping through meta-features. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, pages 119–128. ACM, 2013. *Cité page 42*

Jun Zhu, Zaiqing Nie, Xiaojiang Liu, Bo Zhang, and Ji-Rong Wen. Statsnowball : a statistical approach to extracting entity relationships. In *Proc. of WWW 2009, Madrid, Spain*, pages 101–110, 2009. *Cité page 42*

Arkaitz Zubiaga, Damiano Spina, Raquel Martinez, and Victor Fresno. Real-time classification of twitter trends. *Journal of the Association for Information Science and Technology*, 66(3) :462–473, 2015. https://doi.org/10.1002/asi.23186. *Cité page 29*

It's a poor sort of memory that only works backwards

—White Queen

ANNEXE A

Collections de tests utilisées

Campagne d'évaluation	Année	Tâche	Documents	Requêtes	Références	Dans ce mémoire	Publications associées
INEX	2003	Adhoc Content-only	∼ 12000 documents provenant de 21 revues IEEE Computer Society	36 requêtes CO	(Fuhr and Lalmas, 2004)	Section 1.2.1.1	(Sauvagnat et al., 2006c)
INEX	2003	Adhoc Strict Content-And-Structure	∼ 12000 documents provenant de 21 revues IEEE Computer Society	30 requêtes CAS	(Fuhr and Lalmas, 2004)	Section 1.2.1.1	(Sauvagnat et al., 2006b)
INEX	2004	Adhoc Content-only	∼ 12000 documents provenant de 21 revues IEEE Computer Society	39 requêtes CO	(Malik et al., 2005)	Section 1.2.1.1	(Sauvagnat et al., 2006c)
INEX	2005	Adhoc Content-only	∼ 17000 documents provenant de 21 revues IEEE Computer Society	28 requêtes CO	(Malik et al., 2006)	Section 1.2.1.1	(Pinel-Sauvagnat and Boughanem, 2006, 2007; Sauvagnat and Boughanem, 2006; Sauvagnat et al., 2005, 2006a)
INEX	2005	Adhoc Content-only + Structure	∼ 17000 documents provenant de 21 revues IEEE Computer Society	28 requêtes CO+S	(Malik et al., 2006)	Section 1.2.1.1	(Sauvagnat et al., 2005, 2006a)
INEX	2005	Relevance Feedback	∼ 17000 documents provenant de 21 revues IEEE Computer Society	28 requêtes CO / CO+S, 12 requêtes VVCAS	(Malik et al., 2006)	Section 1.2.1.3	(Hlaoua and Pinel-Sauvagnat, 2006; Hlaoua et al., 2006a,b, 2007a,c,d,d, 2010; Sauvagnat et al., 2005)
INEX	2006	Relevance Feedback	∼ 660 000 documents issus de Wikipedia	114 requêtes CO+S	(Malik et al., 2007)	Section 1.2.1.3	(Hlaoua et al., 2007b,e, 2010)

TABLEAU A.1 – Collections de test pour la recherche adhoc et la la réinjection de pertinence - Recherche de granules XML

Campagne d'évaluation	Année	Tâche	Documents	Requêtes	Références	Dans ce mémoire	Publications associées
INEX	2005	Strict Content and Structure (SSCAS)	∼ 17000 documents provenant de 21 revues IEEE Computer Society	8 requêtes	(Malik et al., 2006)	Section 1.2.1.2	(Laitang and Pinel-Sauvagnat, 2011; Laitang et al., 2011, 2012b, 2013a)
INEX	2010	Datacentric	IMDb (1.6 millions de films, 1.8 millions d'acteurs,...)	26 requêtes	(Trotman and Wang, 2011)	Section 1.2.1.2	(Laitang et al., 2012b, 2013a,b)
INEX	2011	Datacentric	IMDb (1.6 millions de films, 1.8 millions d'acteurs,...)	38 requêtes	(Wang et al., 2012)	Section 1.2.1.2	(Laitang et al., 2012a)
INEX	2005	Multimedia	462 documents *Lonely Planet*	19 requêtes	(van Zwol et al., 2006)	Section 1.2.2	
INEX	2006	Multimedia Fragment	∼ 660 000 documents issus de Wikipedia	9 requêtes	(Westerveld and van Zwol, 2007a)	Section 1.2.2	(Hlaoua et al., 2007e; Torjmen et al., 2007b, 2008a,b, 2009c, 2010; Torjmen-Khemakhem et al., 2013)
INEX	2006	Multimedia Image	171 900 documents images issus de Wikipedia	9 requêtes	(Westerveld and van Zwol, 2007a)	Section 1.2.2	(Torjmen et al., 2007b)
INEX	2007	Multimedia Fragment	∼ 660 000 documents issus de Wikipedia	19 requêtes	(Tsikrika and Westerveld, 2008)	Section 1.2.2	(Torjmen and Pinel-Sauvagnat, 2009; Torjmen et al., 2007a, 2009c, 2010; Torjmen-Khemakhem et al., 2013)
INEX	2007	Multimedia Image	171 900 documents images issus de Wikipedia	20 requêtes	(Tsikrika and Westerveld, 2008)	Section 1.2.2	(Torjmen et al., 2007a)
CLEF	2007	ImageClef Photographic	IAPR TC-12 Benchmark (20,000 photos en couleur avec légende au format XML)	60 requêtes textuelles	(Grubinger et al., 2008)	Section 1.2.2	(Torjmen et al., 2008e)
CLEF	2007	ImageClef Medical Retrieval	66 0000 radiographies médicales	30 requêtes	(Müller et al., 2007)	Section 1.2.2	(Torjmen et al., 2008e)
CLEF	2008	ImageClef WikipediaMM	171 900 documents images issus de Wikipedia (idem INEX 2007)	75 requêtes	(Tsikrika and Kludas, 2009)	Section 1.2.2	(Torjmen et al., 2009b)
CLEF	2008	ImageClef Medical Retrieval	66 0000 radiographies médicales	30 requêtes	(Müller et al., 2008)	Section 1.2.2	(Torjmen et al., 2009a)

TABLEAU A.2 – Collections de test pour la recherche orientée structure et pour la recherche multimedia - Recherche de granules XML

Campagne d'évaluation	Année	Tâche	Documents	Requêtes	Références	Dans ce mémoire	Publications associées
TREC	2011	Microblog	16 millions de tweets	49 topics	(Ounis et al., 2011)	Section 2.2.2	(Damak et al., 2011, 2012, 2013)
TREC	2012	Microblog	16 millions de tweets	60 topics	(Ounis et al., 2012)	Section 2.2.2	(Ben Jabeur et al., 2012a; Damak et al., 2013)
TREC	2013	Microblog	240 millions de tweets	60 topics	(Lin and Efron, 2013)	Section 2.2.1	(Ben Jabeur et al., 2013)
TREC	2015	Microblog	Flux temps-réel sur 10 jours	50 profils	(Lin et al., 2015)	Section 2.2.3	(Chellal et al., 2015; Palmer et al., 2017)
TREC	2016	Real-Time Summarization	Flux temps-réel sur 10 jours	50 profils	(Lin et al., 2016)	Sections 2.2.3, 5.2.4	(Hubert et al., 2017b; Moulahi et al., 2016)
TREC	2017	Real-Time Summarization	Flux temps-réel sur 10 jours	50 profils	(Lin et al., 2016)	Sections 2.2.3, 5.2.4	(Hubert et al., 2017a,b)

TABLEAU A.3 – Collections de test pour la recherche de microblogs

Campagne d'évaluation	Année	Tâche	Documents	Requêtes	Références	Dans ce mémoire	Publications associées
TREC	2013	KBA (Knowledge Base Acceleration) - CCR (Cumulative Citation Recommandation	1 milliard de documents Web crawlés d'octobre 2011 à février 2013	141 profils	(Frank et al., 2013)	Section 3.2.2	(Abbes et al., 2013b, 2014c, 2015d,e)
TREC	2014	KBA (Knowledge Base Acceleration) - CCR (Cumulative Citation Recommandation	1,2 milliard de documents (KBA 2013 + crawl jusqu'à mai 2013)	109 profils	(Frank et al., 2014)	Section 3.2.2	(Abbes et al., 2014b)
TREC	2014	Temporal Summarization	Corpus KBA 2013	9 topics	(Aslam et al., 2013a)	Section 4.2.2	(Abbes et al., 2014a, 2015a,b)
TREC	2015	Temporal Summarization	Corpus KBA 2013 filtré	15 topics	(Aslam et al., 2015)	Section 4.2.2	(Abbes et al., 2015c)

TABLEAU A.4 – Collections de test pour la recherche de documents et de phrases vitales

ANNEXE B

Curriculum vitae

Karen Pinel-Sauvagnat

Unité de Recherche : IRIT (Institut de Recherche en Informatique de Toulouse)
Équipe : IRIS (Information Retrieval and Information Synthesis)
Établissement d'affectation : Université Paul Sabatier, Toulouse 3
Section CNU : 27

Adresse : IRIT,
Equipe IRIS,
118 route de Narbonne
31042 Toulouse Cedex 9
France

Téléphone : +33 5 61 55 63 22
Mail : karen.sauvagnat@irit.fr
Web : http://www.irit.fr/-Karen.Pinel-Sauvagnat/

1. Déroulement de la carrière

Depuis 2006 — Maître de Conférences, Université Paul Sabatier, Toulouse 3.
Congés maternité en 2007, 2009 et 2015

2005-2006 — Attachée Temporaire d'Enseignement et de Recherche(1/2 poste),
Université des Sciences Sociales (maintenant Université Toulouse Capitole), Toulouse 1

2. Formation universitaire

Sept. 2002 - Juin 2005 — Doctorat en informatique de l'Université Paul Sabatier de Toulouse
Modèle flexible pour la recherche d'information dans des corpus de documents semi-structurés [pdf]
Au sein de l'équipe SIG (Systèmes d'Information Généralisés) de l'IRIT sous la direction de Mohand Boughanem et Claude Chrisment

2002 — DEA DISIC (Documents multimédias, Images, et Systèmes d'Information Communicants), INSA de Lyon
Mention Bien (rang : 2/30)
Conception de systèmes intelligents pour la télémédecine citoyenne en cardiologie
Au sein de l'unité INSERM ERM107, Hôpital de Bron, sous la direction de Paul Rubel

1999-2002 — Diplôme d'ingénieur en Informatique de l'INSA de Lyon
Félicitations du jury (rang : 2/120)

2000-2001 — Année d'échange Erasmus à Karlsruhe (Allemagne)

1997-1999 — DEUG MIAS (Mathématiques, Informatique Appliqués aux Sciences)
Université Blaise Pascal, Clermont-Ferrand
Mention Très Bien (Major de promotion)

3. Activité scientifique

3.1. Thématique de recherche

Mes recherches se situent dans la thématique de la Recherche d'Information (RI). Plus précisément, je m'intéresse à :

- la recherche d'information agrégée
  - Recherche agrégée relationnelle : recherche d'attributs, de relations, etc.
  - Recherche agrégée inter-verticale
  - Filtrage et agrégation d'information vitales autour des entités
- la recherche d'information dans des microblogs (notamment recherche d'information temps-réel)
- l'évaluation de la recherche d'information : protocoles, mesures de pertinences, etc.

Je me suis, par le passé, beaucoup intéressée à la RI dans des corpus de documents semi-structurés (de type XML) : recherche adhoc, utilisation de la théorie des graphes pour la recherche, réinjection de pertinence, et recherche d'images.

3.2. Publications scientifiques

La liste complète de mes publications depuis la fin de ma thèse, c'est-à-dire depuis Juillet 2005, est présente en Annexe. Cette liste est également disponible sur le serveur de l'IRIT. Les mentions « Poster » et « Short paper » sont rajoutées à la fin de la référence des articles de ce type. Lorsque connus, les indicateurs propres à chaque publication sont fournis: (i) pour les journaux : IF, IF5 : Impact Facteur et Impact Facteur à 5 ans; classement selon SJR et (ii) pour les conférences : AR (Acceptance Rate), Core. Les indicateurs sont reportés sur l'année correspondant à la publication, ou l'année la plus proche disponible le cas échéant.
Profil DBLP
Profil Google Scholar
ORCID

Le tableau suivant résume ces publications par catégorie et par audience.

Revues		**Conférences**		**Actes**		**Chapitres d'ouvrage**		**Total**	
Int.	Nat.	Int.	Nat.	Int.	Nat.	Int.	Nat.	Int.	Nat.
8	3	40†	17*	1	/	1(+1 révision)	1	49	21

†: Dont 9 courts ou posters, et 18 liés à des campagnes d'évaluation
*: Dont 4 courts

Mes publications à caractère pédagogique (2 ouvrages chez Hermes [P1], [P3] et 2 articles dans la revue Techniques de l'Ingénieur [P2], [P4]), ne figurent pas dans le tableau.

Les cinq publications qui me paraissent les plus significatives et illustrant le mieux la diversité de mes travaux sont, en les hiérarchisant, les publications [RI8] , [RI6], [CI13], [RI5], et [RI4]. Les publications [RI6], [CI13], concernent la recherche d'information agrégée, les [RI5], [RI4] la recherche d'information structurée. La publication [RI8] illustre mes recherches actuelles.

3.3. Conception et développement de prototypes logiciels

2014-2015 Application Vizurbi (https://www.irit.fr/ Karen.Pinel-Sauvagnat/vizurbi/Toulouse.php pour la version de démonstration, fonctionne sous Google Chrome version >=37). Cette application Web propose une visualisation isochrone des temps de transports en commun, et est basée sur des Open Data libérées par la ville de Toulouse et la société Tisseo. Elle a été développée en partenariat avec Aurélien Garivier, professeur à l'Institut de Mathématiques de Toulouse. L'application PHP/Javascript est composée d'environ 1800 lignes de code, sous licence GNU General Public License. Elle nous a permis de participer au concours Open Data 2014 de la ville de Toulouse, et a été utilisée lors d'un projet d'envergure au sein de la formation SID de l'Université Paul Sabatier, afin de faire travailler les étudiants des 3 promotions (environ 90 étudiants) sur son amélioration.

2013 Développement d'un prototype pour participer au challenge Yandex 2013 ("Personalized Web Search Challenge"), en collaboration avec Gilles Hubert et Guillaume Cabanac. Il s'agissait de personnaliser des résultats de recherche à partir de logs fournis aux participants. Notre prototype nous a permis de nous classer au rang 37 sur 194 participants.

2004 - 2009 Prototype XFIRM : moteur de recherche XML, conçu et développé durant ma thèse, et augmenté de modules spécifiques par la suite.

Participation à des campagnes d'évaluation internationales

Les campagnes internationales offrent un cadre international pour la confrontation d'approches et de systèmes de recherche d'information. La participation à ces campagnes demande un investissement important : réflexion sur les approches à tester, développement des algorithmes sur des collections souvent très volumineuses (pouvant atteindre plusieurs To), envoi des résultats selon un calendrier fixé. Ces campagnes sont les campagnes TREC (Text Retrieval Conference -http://trec.nist.gov), INEX (Initiative for the Evaluation of XML Retrieval - http://inex.mmci.uni-saarland.de/) et CLEF (Cross-Language Evaluation Forum- www.clef-initiative.eu). Chaque participation à une campagne a donné lieu à un article de compte-rendu publié dans les actes des conférences adossées aux campagnes.

2017 TREC Real Time Summarization [CE18]
2016 TREC Real Time Summarization [CE17]
2015 TREC Microblog [CE16] • TREC Temporal Summarization [CE15]
2014 TREC KBA [CE14] • TREC Temporal Summarization [CE13]
2013 TREC Microblog [CE11] • TREC Contextual [CE10]• TREC KBA (Knowledge Base Acceleration) [CE12]
2012 TREC Microblog [CE9]
2011 TREC Microblog [CE7] • INEX Data centric track [CE8]
2008 ImageCLEF [CE5] [CE6]
2007 INEX Multimedia track [CE4] • ImageCLEF [CE3]
2006 INEX ad-hoc track [CE2] • INEX Relevance feedback track [CE2] • INEX Multimedia track [CE2]
2005 INEX ad-hoc track [CE1] • INEX Relevance feedback track [CE1]

3.4. Encadrement et formation à la recherche

Étudiants en thèse

Sept. 2014 - Nov. 2017 — Thibaut Thonet, *Modèles thématiques pour la découverte non supervisée de points de vue sur le Web* [pdf]
(avec Guillaume Cabanac (directeur) et Mohand Boughanem)
Financement : contrat doctoral
Taux d'encadrement : 30%
Publications communes : [CI22], [CI21]
Devenir de l'étudiant : Post-doc au LIG (Grenoble)

Sept. 2014 - Avr. 2017 — Thomas Palmer, *Recherche d'information contextuelle dans des flux de données*
(avec Gilles Hubert)
Financement: sur projet FUI ACOVAS
Thèse suspendue en Avril 2017.
Taux d'encadrement : 50%
Publications communes : [CE16], [CE17], [C17]
Devenir de l'étudiant : Chef SI, Patrimoine SA Languedocienne (Toulouse)

Sept. 2012 - Déc. 2015 — Rafik Abbes, *Filtrage et agrégation d'informations vitales relatives à des entités* [pdf]
(avec Nathalie Hernandez et Mohand Boughanem)
Financement : bourse tunisienne
Taux d'encadrement : 30%
Publications communes : [RI7], [CI20], [CI19] , [CE15], [CE14], [CE13], [CE12], [C16], [C15], [C14]
Devenir de l'étudiant : Maître assistant à l'Institut Supérieur d'Informatique et de Multimédia de Gabès (Tunisie)

Sept. 2010 - Juin 2014 — Firas Damak, *Étude des facteurs de pertinence dans la recherche de microblogs* [pdf]
(avec Guillaume Cabanac et Mohand Boughanem)
Financement : sur projet européen QUAERO
Taux d'encadrement : 30%
Publications communes : [CI16], [CE11], [CE9], [CE7], [C13]
Devenir de l'étudiant : Spécialiste « data science » dans une société d'informatique à Londres (Royaume-Uni)

Sept. 2009 - Juil. 2013 — Cyril Laitang, *Impact de la structure des documents XML sur le processus d'appariement dans le contexte de la Recherche d'Information Semi-struturée* [pdf]
(avec Mohand Boughanem)
Financement : sur projet ANR AOC
Taux d'encadrement : 50%
Publications communes : [RI4], [CI18], [CI17], [CE8], [C12], [C9]
Devenir de l'étudiant : Ingénieur Informatique, Canada

Sept. 2008 - Dec. 2011 — Arlind Kopliku, *Approaches to implement and evaluate aggregated search* [pdf]
(avec Mohand Boughanem)
Financement : contrat doctoral
Taux d'encadrement : 50%
Publications communes : [RI6], [CI14], [CI13], [CI12], [CI11], [R2], [C11], [C10] , [C7]

Devenir de l'étudiant: Spécialiste Big Data free-lance pour Airbus (Toulouse)

Sept. 2006 - Nov. 2009
Mouna Torjmen, *Approches de Recherche Multimedia dans des Documents Semi-Structurés : Utilisation du contexte textuel et structurel pour la sélection d'objets multimedia* [pdf]
(avec Mohand Boughanem)
Financement : sur projet
Taux d'encadrement : 50%
Publications communes : [RI2], [CI10], [CI9], [CE6], [CE5], [CE5], [CE3],[C6], [C5], [C4]
Devenir de l'étudiant : Maître assistante à l'Université de Sfax (Tunisie)

Sept. 2004 - Dec. 2007
Lobna Hlaoua, *Refomulation de requêtes par réinjection de pertinence dans les documents semi-structurés* [pdf]
(avec Mohand Boughanem)
Taux d'encadrement : 50%
Publications communes : [RI3], [CI8], [CI6], [CI5], [CI4], [CI3], [CE2], [CE1],[C3], [C2]
Devenir de l'étudiant : Maître assistante à l'Université de Sousse (Tunisie)

Étudiants en Master

2014
Thibaut Thonet, étudiant en troisième année du cycle ingénieur ENSEEIHT
Recherche d'information agrégée et aide à la lecture de news
(avec Mohand Boughanem et Guillaume Cabanac)

2012
Rafik Abbes, étudiant en Master II Recherche
Étude de l'apport du Web de données et du Web relationnel dans la recherche agrégée relationnelle

2010
Firas Damak, étudiant en Master II Recherche
Étude de l'impact de l'agrégation de recherches verticales sur la qualité des SRI

2010
Ines Krichen, étudiante en Master II Recherche
Extraction, sélection et agrégation d'information à partir d'une base de connaissance

2009
Mai Dong Le, étudiant en Master II Recherche
Graphes et appariement sémantique de documents XML

2008
Julien Boujot, étudiant en Master II Recherche
Utilisation des liens pour la recherche dans les documents structurés : XMLRank

2006
Mouna Torjmen, étudiante en Master II Recherche
Recherche contextuelle d'images dans les documents XML

2004
Lobna Hlaoua, étudiante en DEA
Recherche d'Information dans des Documents XML : Utilisation d'une Technique de Propagation de la Pertinence

3.5. Rayonnement

3.5.1. Expertise de projets

2018 Évaluatrice pour l'ANR (Appel à projet générique, catégorie « La Révolution numérique : rapports au savoir et à la culture »)

2017 Évaluatrice pour l'ISF (Israël Science Foundation)

2014, 2015 Évaluatrice externe de projets pour le NSERC (Natural Sciences and Engineering Research Council of Canada)

2014 Évaluatrice pour l'ANR (Appel à projet générique, catégorie « Culture-patrimoine »)

3.5.2. Comités de rédaction / progamme

Comité de rédaction

Depuis 2016 Membre du comité de rédaction de la revue ISTE Open Science « Recherche d'information, document et web sémantique », dirigée par Vincent Claveau (IRISA-CNRS). https://www.openscience.fr/Recherche-d-information-document-et-web-semantique

Présidences de comités de programme

2015 Co-présidente du comité de programme de la conférence internationale CLEF 2015 (http://clef2015.clef-initiative.eu/ (avec Jaap Kamps)

2013 Présidente du comité de programme des 8e RJCRI (RJCRI 2013 - Rencontres Jeunes Chercheurs en Recherche d'information, http://coria.unine.ch/rjcri.htm), Neuchâtel, Suisse

2011 Co-présidente du comité de programme et d'organisation du Workshop GAOC (GAOC 2011 - http://www.irit.fr/GAOC2011), Brest, France

Participation à des comités de programme

Les conférences nationales apparaissent en orange, les internationales en vert.

2018 ECIR • CORIA • Atelier RISE

2017 SIGIR (papiers courts) • ECIR • CORIA • CLEF

2016 SIGIR (papiers courts) • ECIR • CORIA • CLEF • CIKM • KDIR

2015 ECIR • CORIA

2014 SIGIR (papiers courts) • ECIR • CORIA

2013 Symposium SIIM • Atelier RISE • CORIA

2012 CORIA

2011 CIKM • Symposium SIIM • SIGIR (posters)

2005 SIGIR (posters)

Sigles :

- CIKM : Conference on Information and Knowledge Management
- CLEF : Conference and Labs for the Evaluation Forum
- CORIA : COnférence en Recherche d'Information et Applications
- ECIR : European Conference for Information Retrieval
- RISE : Recherche d'Information SEmantique
- SIIM : Symposium sur l'Ingénierie de l'Information Médicale
- SIGIR : ACM SIGIR Conference on Research and Development in Information Retrieval

Relectrice pour les revues

2016 JASIST
2015 IR
2014 TOIS • Document Numérique
2013 TKDE • I3
2011 IPM • IR • I3
2010 IPM
2006 IR

Sigles:

- I3 : Interaction Intelligence Information
- IR : Information Retrieval Journal
- IPM : Information Processing and Management
- JASIST : Journal of the Association for Information Science and Technology
- TKDE : Transactions on Knowledge and Data Engineering - IEEE
- TOIS : Transactions on Information Systems - ACM

Relectrice additionnelle

2013 SAC
2011 WI •COSI
2010 RIAO • CIKM • CORIA
2009 ECIR • EGC • Inforsid
2008 ICEIS • SIGIR • EGC • ECIR • SAC
2007 ICEIS • RIAO • SAC • CIKM • ECIR • RCIS • SIGIR
2006 CORIA • SIGIR • FQAS • SAC • SPIRE
2005 CIKM

3.5.3. Comités d'organisation

2014, 2016 Organisatrice, avec Ludovic Denoyer (LIP6), des CORIA-CIFED Hackdays 2014 et 2016. Ces deux évènements ont réuni entre 25 et 30 participants, qui ont travaillé pendant 24h sur des prototypes et applications vitrines de la communauté de Recherche d'Information, autour d'un thème « imposé » (Computational Cooking en 2014, Séries télévisées en 2016).
(http://hackday.lip6.fr)
Les résultats et prototypes ont ensuite été présentés en conférence plénière.

Participation à des comités d'organisation

2016 Semaine du Document Numérique et de la Recherche d'Information (Conférences CORIA et CIFED), Toulouse, France -http://www.irit.fr/sdnri2016/
2015 CLEF (Cross-Language Evaluation Forum) 2015 Conference, Toulouse, France - http://clef2015.clef-initiative.eu/
2013 Conférence Extraction and Gestion des Connaissances, Toulouse, France - http://www.irit.fr/EGC2013
2011 Symposium SIIM (Ingénierie de l'Information Médicale), Toulouse, France - http://www.irit.fr/SIIM
2010 Conférence BDA (Bases de Données Avancées), Toulouse, France - http://www.irit.fr/BDA2010

2009 Conférence Inforsid, Toulouse, France - http://www.irit.fr/inforsid2009
European Conference on Information Retrieval (ECIR), Toulouse, France - http://ecir09.irit.fr

2008 Ecole d'automne en Recherche d'Information et Applications (EARIA), Boussens, France

3.5.4. Animation scientifique

Depuis 2015 Chargée de communication au sein de l'ARIA (Association pour la Recherche d'Information et Applications, http://www.asso-aria.org/)

Octobre 2016 Organisation des EARIA Days (avec Benjamin Piwowarski, LIP 6), Ecole d'automne EARIA (École d'Automne en Recherche d'Information et Applications - http://www.asso-aria.org/index.php?option=com_content&view=article&id=142&Itemid=540).

Mai 2015 Présentation « Recherche d'information, un rapide panorama » durant la journée organisée à l'RIT pour les enseignants en lycée de l'option ISN (Informatique et Sciences du Numérique).

Octobre 2014 Organisation de la soirée Hack Evening (avec Ludovic Denoyer, LIP6) , École d'automne EARIA (École d'Automne en Recherche d'Information et Applications - http://www.asso-aria.org/index.php?option=com_content&view=article&id=108&Itemid=496).

3.5.5. Prix & distinctions

2015 Prix du meilleur article de la Conférence SAC (ACM Symposium on Applied Computing , catégorie Information Systems [CI19]

2014-2017 PEDR – rang A

2011-2014 Prime d'excellence scientifique – rang B

3.6. Participation à des projets

2015-2018 Projet ANR CAIR (Contextual Aggregated Information Retrieval - http://www.irit.fr/CAIR)
Partenaires : LAMSADE (Paris Dauphine), PRISM (Versailles), LIRIS (Lyon), Telecom Sud Paris. L'IRIT est coordinateur du projet
Rôle dans le projet : encadrement de thèse, participation à la recherche et à l'élaboration des livrables

2014-2017 Projet FUI ACOVAS (Outils Agile pour la COnception et VAlidation Système)
Partenaires: Airbus, GFI, Liebherr, Nexeya, Prometil, Zodiac Aerospace
Rôle dans le projet : encadrement de thèse

2013-2014 **Coordinatrice** du projet PEPS INS2I RICA (Recherche d'Information en Contexte et Agrégée - http://www.irit.fr/RICA)
Partenaires : LAMSADE (Paris Dauphine), PRISM (Versailles), LIRIS (Lyon), Telecom Sud Paris

2009-2012 Projet ANR AOC (Appariement d'Objets Complexes, http://aoc.irit.fr)
Partenaires : PRISM (Versailles), IRISA (Lannion), LIRIS (Lyon), LIESP (Lyon). L'IRIT est coordinateur du projet

2008 - 2012 Projet européen QUAERO (http://www.quaero.org/)
Responsable de Workpackage, Mise en place de plusieurs campagnes d'évaluation, pour l'indexation multimedia ou la recherche d'information sur le Web.

4. Activités pédagogiques

4.1. Thématiques d'enseignements et enseignements effectués

J'effectue mes enseignements au sein de l'Université Paul Sabatier, principalement dans la thématique des **Bases de Données** (bases de données relationnelles, bases de données semi-structurées/XML, mapping objet/relationnel). J'interviens actuellement dans diverses formations :

- SID - Statistique et Informatique Décisionnelle
- STRI - Systèmes de Télécommunications, Réseaux et Télécommunications
- MIAGe - Méthodes Informatiques Appliquées à la Gestion
- DC - Données et Connaissances
- ILORD - Ingénierie du Logiciel, des Réseaux et des Systèmes Distribués

J'ai effectué jusqu'à 350 heures d'enseignements (ETD) par an avant d'obtenir la PES en 2010, j'effectue aujourd'hui environ 250 heures.
Le tableau suivant donne un aperçu de ma charge actuelle : l'offre de formation ayant beaucoup évolué à l'université Paul Sabatier sur l'habilitation 2016-2020, un grand nombre d'UE sont nouvelles. Les tableaux suivants mettent en avant certaines UE dont j'ai été responsable dans le passé, ou effectuées dans des écoles/universités autres que l'Université Paul Sabatier. Il ne s'agit pas d'un descriptif exhaustif de tous les enseignements auxquels j'ai participé.

Les matières/UE dont je suis/ai été responsable sont suivies de la marque †. La responsabilité d'UE implique la préparation des Cours, TD, TP, le recrutement et la gestion des éventuels autres intervenants, la gestion des examens de 1ère et seconde session ainsi que la participation aux jurys de formations. Les variations de volumes horaires au sein d'une même UE sont dues aux effectifs changeants des formations, et au fait que sur certaines années, j'ai confié la réalisation de certains TPs à des Doctorants Chargés d'Enseignement ou à des ATER.

À ces UE s'ajoutent le suivi d'étudiants en stage dans les diverses formations, ainsi que le suivi d'alternants des formations MIAGE, STRI et SID (un ou 2 étudiants par an). Dans ce dernier cas, le suivi nécessite des visites régulières avec l'entreprise et l'étudiant, la participation à des réunions de concertation, ainsi que la correction de rapports et sujets de réflexion.

Charge actuelle					
Matières	Formation	Niveau (Sem.)	Type (Effectif)	Vol. Horaire (ETD)	Années
Transformation de modèles (†)	DC	M2 (S9)	C/TD/TP (15)	~36h/an	Depuis 2016
Big Data et NoSQL (†)	ILORD	M2 (S9)	C/TP (15)	~20h/an	Depuis 2016
Projet Open Data(†) ‡	SID	L3-M1-M2	Projet (80)	-	2014, et depuis 2016
Bases de données avancées (†)	STRI	M1 (S7)	C/TD/TP (60)	~36h/an	Depuis 2006
Projet Base de Données et Web (†)	SID	L3 (S6)	Projet (35)	~35h/an	Depuis 2016
Représentation de données (†)	MIAGE et e-MIAGE	L3 (S5)	C/TD/TP (50)	~50h/an	Depuis 2014 (et 2008 pour e-MIAGe)
Normalisation, accès concurrents et mise en oeuvre de BD	MIAGE	L3 (S6)	TP (50)	~15h/an	Depuis 2014
Bases de données réparties	MIAGE	M1 (S7)	TP (50)	~15h/an	Depuis 2016
Bases de données Web	MIAGE	M1 (S8)	TP (50)	~15h/an	Depuis 2016

Anciennes UE notables					
Matières	Formation	Niveau (Sem.)	Type (Effectif)	Vol. Horaire (ETD)	Années
Langages de requêtes (†)	SID	L3 (S5)	C/TD/TP (20-40)	Entre 30 et 60 h/an	2006-2016
Organisations semi-structurées (†)	MIAGE	M2 (S10)	C/TD/TP (50)	~50h/an	2008-2016
Bases de données structurées et semi-structurées (†)	ASIC	M2 (S9)	C/TD/TP (30)	~36h/an	2011-2016
Médiation de sources de données hétérogènes	Master recherche	M2 (S9)	C (15)	10h/an	2010-2014
Bases de données (†) *	Licence Informatique	L2 (S4)	C-TD/TP **(120-140)**	Environ 60h/an	2011-2013
Systèmes d'information et applications web (†)	Licence informatique	L2 (S4)	C-TD/TP (60)	~50h/an	2010-2012

‡: En 2014, pour sa première édition, l'UE a été mise en place autour de l'application Vizurbi (https://www.irit.fr/ Karen.Pinel-Sauvagnat/vizurbi/Toulouse.php pour la version de démonstration), développée avec Aurélien Garivier de l'IMT autour de données de transports en commun libérées par la régie Tisseo. Nous avons fait collaborer les 3 promotions (80 étudiants) pendant 15 jours à plein temps dans le but d'améliorer l'application et de la présenter à Tisseo. Ceci a nécessité un travail en amont important de découpage des tâches, puis de suivi de projet durant les 15 jours.
L'UE a actuellement pour objectif de faire travailler les 3 promotions de la formation SID autour d'un gros projet lié aux données Open Data et mettant en oeuvre des compétences en Informa-

tique et Statistique. Je coordonne l'UE et les enseignants encadrant le projet (nouvelle équipe tous les ans).
* : Cette UE de licence a demandé une très forte implication: mise en place de tous les supports de cours/TD/TP à partir de zéro, gestion des intervenants (4 à 6 groupes de Cours-TD , 8 groupes de TP), coordination de la correction des projets et de l'examen terminal.

Interventions hors Université Paul Sabatier					
Matières	Formation	Niveau (Sem.)	Type (Effectif)	Vol. Horaire (ETD)	Années
Données hétérogènes	ASIC - Université Toulouse 1	M2 (S10)	C/TD (30)	De 10 à 20h/an	Depuis 2013
Représentation des connaissances et accès à l'information	INSA	5ème année Informatique (5 IF)	C/TP (30)	Entre 10 et 30h/an	2007-2013
Bases de données	ENSEEIHT	3eme année, Dept. GEA	C/TP (25)	10h/an	2007-2009

4.2. Publications pédagogiques

J'ai participé à la rédaction de deux ouvrages sur les bases de données [P1][P3] parus aux éditions Hermes, ainsi que deux articles dans la revue Techniques de l'Ingénieur [P2][P4].

4.3. Responsabilités pédagogiques

2007-2013 Co-responsabilité côté informatique du L3 SID (Système d'Information Décisionnels), formation bi-disciplinaire Mathématiques/Informatique. Cette responsabilité impliquait la construction et le suivi des emplois du temps, le suivi des élèves, la coordination pédagogique des enseignements, leur évaluation, etc.

5. Responsabilités collectives

Participation aux conseils

2018-2021 Membre nommée du Collège Scientifique Informatique de l'Université Paul Sabatier
Membre élue du Groupe d'Avancement Corps B

Participation à des comités de sélection

2018 Poste de Maître de Conférences à l'ESPE de Marseille
2014 Poste de Maître de Conférences 0130 à l'INSA de Lyon
2013 Poste de Maître de Conférences 1720 à l'Enssat, Lannion

Annexe: Publications

Publications internationales

Édition d'ouvrages

[OI1] Josiane Mothe, Jacques Savoy, Jaap Kamps, Karen Pinel-Sauvagnat, Gareth J.F. Jones, Eric San Juan, Linda Cappellato, and Nicola Ferro, editors. *Experimental IR Meets Multilinguality, Multimodality, and Interaction. 6th International Conference of the CLEF Association, CLEF'15, September 8-11, 2015, Proceedings, Toulouse, France, 08/09/2015 - 11/09/2015*, http://www.springerlink.com, septembre 2015. Springer. http://www.springer.com/fr/book/9783319240268, http://doi.org/10.1007/978-3-319-24027-5.

Chapitres d'ouvrages

[ChI2] Karen Pinel-Sauvagnat. Propagation-Based Structured Text Retrieval. In Ling Liu and M. Tamer Özsu, editors, *Encyclopedia of Database Systems, revised edition*. Springer, août 2017. https://doi.org/10.1007/978-1-4899-7993-3_281-2, Sur invitation.

[ChI1] Karen Pinel-Sauvagnat. Propagation-based structured text retrieval. In Ling Liu and M. Tamer Özsu, editors, *Encyclopedia of Database Systems*, pages 2197–2201. Springer, mai 2009. http://dx.doi.org/10.1007/978-0-387-39940-9_281. Sur invitation.

Revues internationales

[RI8] Gilles Hubert, Yoann Pitarch, Karen Pinel-Sauvagnat, Ronan Tournier, and Léa Laporte. TournaRank: When Retrieval Becomes Document Competition. *Information Processing & Management*, 54(2):252–272, mars 2018. https://doi.org/10.1016/j.ipm.2017.11.006, IF :2.391, SJR: Q1-Q2.

[RI7] Rafik Abbes, Karen Pinel-Sauvagnat, Nathalie Hernandez, and Mohand Boughanem. When temporal expressions help to detect vital documents related to an entity. *Applied Computing Review*, 15(3):49–58, septembre 2015. http://doi.org/10.1145/2835260.2835263.

[RI6] Arlind Kopliku, Karen Pinel-Sauvagnat, and Mohand Boughanem. Aggregated search: a new information retrieval paradigm. *ACM Computing Surveys*, 46(3):1–31, janvier 2014. http://doi.acm.org/10.1145/2523817, IF5 : 7.854, IF : 4.043, SJR: Q1.

[RI5] Mouna Torjmen-Khemakhem, Karen Pinel-Sauvagnat, and Mohand Boughanem. Investigating the document structure as a source of evidence for multimedia fragment retrieval. *Information Processing & Management*, 49:1281–1300, juillet 2013. http://dx.doi.org/10.1016/j.ipm.2013.06.001, IF5 :1.48, IF :1.069, SJR: Q1-Q2.

[RI4] Mohammed Amin Tahraoui, Karen Pinel-Sauvagnat, Cyril Laitang, Mohand Boughanem, Hamamache Kheddouci, and Lei Ning. A survey on tree matching and XML retrieval. *Computer Science Review*, 8:1–23, mai 2013. http://dx.doi.org/10.1016/j.cosrev.2013.02.001, SJR : Q1.

[RI3] Lobna Hlaoua, Karen Pinel-Sauvagnat, and Mohand Boughanem. Relevance Feedback Revisited: Dealing with Content and Structure in XML Documents. *International Journal on Digital Libraries*, 11(1):1–24, mars 2010. http://dx.doi.org/10.1007/s00799-010-0061-5, SJR: Q1.

[RI2] Mouna Torjmen, Karen Pinel-Sauvagnat, and Mohand Boughanem. Using textual and structural context for searching multimedia elements. *International Journal of Business Intelligence and Data Mining, Special Issue on Beyond Multimedia and XML Streams Querying and Mining*, 5(4):323–352, octobre 2010. http://dx.doi.org/10.1504/IJBIDM.2010.036123, SJR: Q2-Q4.

[RI1] Karen Sauvagnat, Mohand Boughanem, and Claude Chrisment. Answering content-and-structure-based queries on XML documents using relevance propagation. *Information Systems, Special Issue SPIRE 2004*, 31:621–635, janvier 2006. http://dx.doi.org/10.1016/j.is.2005.11.007, IF: 1,887, SJR: Q1.

Conférences internationales avec comités de sélection

[CI22] Thibaut Thonet, Guillaume Cabanac, Mohand Boughanem, and Karen Pinel-Sauvagnat. Users Are Known by the Company They Keep: Topic Models for Viewpoint Discovery in Social Networks . In *Conference on Information and Knowledge Management (CIKM), Singapore*, pages 87–96, novembre 2017. ACM. https://doi.org/10.1145/3132847.3132897, AR (171/820, 21%), Core A.

[CI21] Thibaut Thonet, Guillaume Cabanac, Mohand Boughanem, and Karen Pinel-Sauvagnat. VODUM: a Topic Model Unifying Viewpoint, Topic and Opinion Discovery . In *European Conference on Information Retrieval (ECIR), Padua, Italy*, volume 9626 of *LNCS*, pages 533–545, mars 2016. Springer. http://doi.org/10.1007/978-3-319-30671-1_39, AR (42/200, 21%), Core B.

[CI20] Rafik Abbes, Nathalie Hernandez, Karen Pinel-Sauvagnat, and 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, Singapour*, pages 49–58, décembre 2015. IEEE. http://doi.org/10.1109/WI-IAT.2015.32, AR= ((53+46)/173, 56%).

[CI19] Rafik Abbes, Karen Pinel-Sauvagnat, Nathalie Hernandez, and Mohand Boughanem. Leveraging temporal expressions to filter vital documents related to an entity. In *ACM Symposium on Applied Computing (SAC), Salamanca, Spain*, pages 1093–1098, avril 2015. ACM. http://dx.doi.org/10.1145/2695664.2695910, AR =(371/1613), 23%), Core B, Distinction décernée : Best paper award - Information Systems.

[CI18] Cyril Laitang, Karen Pinel-Sauvagnat, and Mohand Boughanem. Estimating Structural Relevance of XML Elements Through Language Model (short paper). In *Open Areas in Information Retrieval (OAIR), Lisbon, Portugal*, mai 2013. Kent State UNiversity. http://oatao.univ-toulouse.fr/12412/, AR= ((16+21)/71, 52%).

[CI17] Cyril Laitang, Karen Pinel-Sauvagnat, and Mohand Boughanem. DTD based costs for Tree-Edit distance in Structured Information Retrieval . In *European Conference on Information Retrieval (ECIR), Moscou, Russie*, pages 158–179. Springer, mars 2013. https://doi.org/10.1007/978-3-642-36973-5_14, AR= (30/191, 29%), Core B.

[CI16] Firas Damak, Karen Pinel-Sauvagnat, Guillaume Cabanac, and Mohand Boughanem. Effectiveness of State-of-the-art Features for Microblog Search . In *ACM Symposium on Applied Computing (SAC), Coimbra, Portugal*, pages 914–919, mars 2013. ACM. http://dx.doi.org/10.1145/2480362.2480537, AR= (255/1063, 24%), Core B.

[CI15] Cyril Laitang, Mohand Boughanem, and Karen Pinel-Sauvagnat. XML Information Retrieval through Tree Edit Distance and Structural Summaries (poster). In *Asia Information Retrieval Society Conference (AIRS), Dubai, United Atab Emirates*, pages 73–83, décembre 2011. Springer. http://dx.doi.org/10.1007/978-3-642-25631-8_7, AR= ((32+29)/132, 46%).

[CI14] Arlind Kopliku, Karen Pinel-Sauvagnat, and Mohand Boughanem. Attribute Retrieval from Relational Web tables . In *Symposium on String Processing and Information Retrieval (SPIRE), Pisa, Italy*, pages 117–128. Springer, octobre 2011. https://doi.org/10.1007/978-3-642-24583-1_12, AR= (30/102, 29%), Core B.

[CI13] Arlind Kopliku, Mohand Boughanem, and Karen Pinel-Sauvagnat. Towards a framework for attribute retrieval . In *Conference on Information and Knowledge Management (CIKM), Glasgow, UK*, pages 515–524, octobre 2011. ACM. https://doi.org/10.1145/2063576.2063654, AR= (137/917, 15%), Core A.

[CI12] Arlind Kopliku, Firas Damak, Karen Pinel-Sauvagnat, and Mohand Boughanem. Interest and Evaluation of Aggregated Search . In *IEEE/WIC/ACM International Conference on Web Intelligence, Lyon*, page (on line),août 2011. ACM. https://doi.org/10.1109/WI-IAT.2011.99, AR= (41/200, 20%).

[CI11] Arlind Kopliku, Karen Pinel-Sauvagnat, and Mohand Boughanem. Retrieving attributes using Web tables (poster). In *Joint Conference on Digital Libraries (JCDL) (JCDL), Ottawa*, pages 397–398, juin 2011. ACM. http://dx.doi.org/10.1145/1998076.1998153, AR= ((28+29+32)/243, 37%), Core A*.

[CI10] Mouna Torjmen, Karen Pinel-Sauvagnat, and Mohand Boughanem. XML Multimedia Retrieval: From relevant textual information to relevant multimedia fragments . In *European Conference on Information Retrieval (ECIR), Toulouse*, pages 150–161, 2009. Springer. http://dx.

doi.org/10.1007/978-3-642-00958-7_16, AR= (42/188, 22%), Core B.

[CI9] Mouna Torjmen, Karen Pinel-Sauvagnat, and Mohand Boughanem. Towards a structure-based multimedia retrieval model (short paper). In *ACM International Conference on Multimedia Information Retrieval , Vancouver, Canada*, pages 350–357, octobre 2008. ACM. http://dx.doi.org/10.1145/1460096.1460153, AR= ((56+12)/308, 21%), Core B.

[CI8] Lobna Hlaoua, Mohand Boughanem, and Karen Pinel-Sauvagnat. Combination of Evidences in Relevance Feedback for XML Retrieval (short paper). In *Conference on Information and Knowledge Management (CIKM), Lisbonne, Portugal*, pages 893–896, novembre 2007. ACM Press. http://dx.doi.org/10.1145/1321440.1321569, AR= ((86+49)/512, 26%, Core A.

[CI7] Karen Pinel-Sauvagnat and Mohand Boughanem. A survey on XML focussed component retrieval . In *Large-Scale Semantic Access to Content (Text, Image, Video and Sound) (RIAO), Pittsburgh, USA*, juin 2007. Centre de hautes études internationales d'Informatique Documentaire (C.I.D.). http://www.irit.fr/~Karen.Pinel-Sauvagnat/fichiers/RIAO2007.pdf, AR= (33/129, 25%).

[CI6] Lobna Hlaoua, Mohand Boughanem, and Karen Pinel-Sauvagnat. Using a Content-and-Structure Oriented Method for Relevance Feedback in XML Retrieval (short paper). In *Large-Scale Semantic Access to Content (Text, Image, Video and Sound) (RIAO), Pittsburgh (PA) États-Unis*, juin 2007. Centre de hautes études internationales d'Informatique Documentaire (C.I.D.). AR ((33+29)/129, 48%)

[CI5] Lobna Hlaoua, Karen Pinel-Sauvagnat, and Mohand Boughanem. Relevance Feedback for XML Retrieval: using structure and content to expand queries. In Colette Rolland, Oscar Pastor, and Jean-Louis Cavarero, editors, *International Conference on Research Challenge in Information Science (RCIS), Ouarzazate- Maroc*, pages 195–202, avril 2007. EMSI - Ecole Marocaine des Sciences de l'Ingénieur.

[CI4] Lobna Hlaoua, Karen Pinel-Sauvagnat, and Mohand Boughanem. A structure-oriented relevance feedback method for XML retrieval (short paper). In *Conference on Information and Knowledge Management (CIKM), Arlington, Virginia, USA.*, pages 780–781, novembre 2006. ACM. http://dx.doi.org/10.1145/1183614.1183727, AR= ((81+56)/537, 25%), Core A.

[CI3] Lobna Hlaoua and Karen Pinel-Sauvagnat. Structure-Oriented Relevance Feedback in XML Retrieval (regular paper). In Vicente P. Guerrero-Note, editor, *International Conference on Multidisciplinary Sciences & Technologies (InSciT), Merida, Espagne*, pages 99–103, octobre 2006. Open Institute of Knowledge.

[CI2] Karen Sauvagnat, Mohand Boughanem, and Claude Chrisment. Why using structural hints in XML retrieval ? . In Henrik Legind Larsen, Gabriella Pasi, and Ortiz-Arroyo Daniel, editors, *Flexible Query Answering (FQAS), Milan, Italie*, Advances in Artificial Intelligence, pages 197–109, juin 2006. Springer. https://doi.org/10.1007/11766254_17.

[CI1] Karen Sauvagnat, Lobna Hlaoua, and Mohand Boughanem. XML retrieval: what about using contextual relevance? (short paper). In *Annual ACM Symposium on Applied Computing (SAC), Dijon*, pages 1114–1120, avril 2006. ACM Press. https://doi.org/10.1007/11766254_17, AR= ((16+5)/54 (IAR), 38%).

Workshops avec comités de sélection

[W1] Damien Palacio, Guillaume Cabanac, Gilles Hubert, Karen Pinel-Sauvagnat, and Christian Sallaberry. Prototyping a Personalized Contextual Retrieval Framework (short paper). In *ACM SIGSPATIAL Workshop on Geographic Information Retrieval (GIR), Orlando, USA*, pages 43–44, novembre 2013. ACM. http://doi.org/10.1145/2533888.2533935.

Participation à des campagnes d'évaluation internationales

Seuls les groupes qui ont répondu (soumis des résultats) à ces campagnes internationales ont le droit de rédiger des articles qui sont ensuite publiés, dans notre cas, soit le site du NIST (TREC), soit dans les actes de la conférence organisée pour la restitution des résultats (INEX, CLEF).

[CE18] Gilles Hubert, José Moreno, Karen Pinel-Sauvagnat, and Yoann Pitarch. Some thoughts from IRIT about the scenario A of the TREC RTS 2016 and 2017 tracks. In *Text REtrieval Conference*

(TREC), Gaithersburg, Maryland USA, November 2017, novembre 2017. National Institute of Standards and Technology (NIST). http://trec.nist.gov/pubs/trec26/papers/IRIT2-RT.pdf.

[CE17] Bilel Moulahi, Lamjed Ben Jabeur, Abdelhamid Chellal, Thomas Palmer, Lynda Tamine, Mohand Boughanem, Karen Pinel-Sauvagnat, and Gilles Hubert. IRIT at TREC Real Time Summarization 2016. In *Text REtrieval Conference (TREC), Gaithersburg, Maryland USA*, novembre 2016. National Institute of Standards and Technology (NIST). http://trec.nist.gov/pubs/trec25/papers/IRIT-RT.pdf.

[CE16] Abdelhamid Chellal, Lamjed Ben Jabeur, Laure Soulier, Bilel Moulahi, Thomas Palmer, Mohand Boughanem, Karen Pinel-Sauvagnat, Lynda Tamine, and Gilles Hubert. IRIT at TREC Microblog 2015 . In *Text REtrieval Conference (TREC), Gaithersburg, Maryland USA*, novembre 2015. National Institute of Standards and Technology (NIST). http://trec.nist.gov/act_part/conference/papers/IRIT-MB.pdf.

[CE15] Rafik Abbes, Bilel Moulahi, Abdelhamid Chellal, Karen Pinel-Sauvagnat, Nathalie Hernandez, Mohand Boughanem, Lynda Tamine, and Sadok Ben Yahia. IRIT at TREC Temporal Summarization 2015. In *Text REtrieval Conference (TREC), Gaithersburg, Maryland USA*, novembre 2015. National Institute of Standards and Technology (NIST). http://trec.nist.gov/act_part/conference/papers/IRIT-TS.pdf.

[CE14] Rafik Abbes, Karen Pinel-Sauvagnat, Nathalie Hernandez, and Mohand Boughanem. IRIT at TREC KBA 2014. In *Text REtrieval Conference (TREC), Gaithersburg, USA*, novembre 2014. National Institute of Standards and Technology (NIST). http://trec.nist.gov/pubs/trec23/papers/pro-IRIT_kba.pdf.

[CE13] Rafik Abbes, Karen Pinel-Sauvagnat, Nathalie Hernandez, and Mohand Boughanem. IRIT at TREC Temporal Summarization 2014. In *Text REtrieval Conference (TREC), Gaithersburg, USA*, novembre 2014. National Institute of Standards and Technology (NIST). http://trec.nist.gov/pubs/trec23/papers/pro-IRIT_ts.pdf.

[CE12] Rafik Abbes, Karen Pinel-Sauvagnat, Nathalie Hernandez, and Mohand Boughanem. IRIT at TREC Knowledge Base Acceleration 2013: Cumulative Citation Recommendation Task. In *Text REtrieval Conference (TREC), Gaithersburg, USA*, novembre 2013. National Institute of Standards and Technology (NIST). http://trec.nist.gov/pubs/trec22/papers/IRIT-kba.pdf.

[CE11] Lamjed Ben Jabeur, Firas Damak, Lynda Tamine, Guillaume Cabanac, Karen Pinel-Sauvagnat, and Mohand Boughanem. IRIT at TREC Microblog Track 2013. In Ellen M. Voorhees, editor, *Text REtrieval Conference (TREC), Gaithersburg, USA*, novembre 2013. National Institute of Standards and Technology (NIST). http://trec.nist.gov/pubs/trec22/papers/IRIT-microblog.pdf.

[CE10] Gilles Hubert, Guillaume Cabanac, Karen Pinel-Sauvagnat, Damien Palacio, and Christian Sallaberry. IRIT, GeoComp, and LIUPPA at the TREC 2013 Contextual Suggestion Track . In Ellen M. Voorhees, editor, *Text REtrieval Conference (TREC), Gaithersburg, USA*, novembre 2013. National Institute of Standards and Technology (NIST). http://trec.nist.gov/pubs/trec22/papers/IRIT-context.pdf.

[CE9] Lamjed Ben Jabeur, Firas Damak, Lynda Tamine, Karen Pinel-Sauvagnat, Guillaume Cabanac, and Mohand Boughanem. IRIT at TREC Microblog 2012: Adhoc Task. In Ellen M. Voorhees and Lori P. Buckland, editors, *Text REtrieval Conference (TREC), Gaithersburg, USA*, novembre 2012. National Institute of Standards and Technology (NIST). http://trec.nist.gov/pubs/trec21/papers/IRIT.microblog.final.pdf.

[CE8] Cyril Laitang, Karen Pinel-Sauvagnat, and Mohand Boughanem. Edit Distance for XML Information Retrieval : Some Experiments on the Datacentric Track of INEX 2011. In *Focused Retrieval of Content and Structure - INEX (Initiative for the Evaluation of XML Retrieval), Imsbach*, Lecture note in computer science, pages 138–145. Springer, avril 2012. https://doi.org/10.1007/978-3-642-35734-3_11.

[CE7] Firas Damak, Lamjed Ben Jabeur, Guillaume Cabanac, Karen Pinel-Sauvagnat, Lynda Tamine, and Mohand Boughanem. IRIT at TREC Microblog 2011. In Voorhees Ellen M. and Buckland Lori P., editors, *Text REtrieval Conference (TREC), Gaithersburg, USA*, novembre 2011. National Institute of Standards and Technology (NIST). http://trec.nist.gov/pubs/trec20/papers/IRIT_SIG.microblog.update.pdf.

[CE6] Mouna Torjmen, Karen Pinel-Sauvagnat, and Mohand Boughanem. Evaluating the impact of image names in context-based retrieval. In *Evaluating Systems for Multilingual and Multimodal Information Access. 9th Workshop of the Cross-Language Evaluation Forum, CLEF 2008, Aarhus, Denmark*, LNCS, pages 756–762, septembre 2009. Springer. https://doi.org/10.1007/978-3-642-04447-2_98.

[CE5] Mouna Torjmen, Karen Pinel-Sauvagnat, and Mohand Boughanem. Methods for combining content-based and textual-based approaches in medical image retrieval (short paper). In *Evaluating Systems for Multilingual and Multimodal Information Access. 9th Workshop of the Cross-Language Evaluation Forum, CLEF 2008, Aarhus, Denmark*, LNCS, pages 691–695, septembre 2009. Springer. https://doi.org/10.1007/978-3-642-04447-2_87.

[CE4] Mouna Torjmen, Karen Pinel-Sauvagnat, and Mohand Boughanem. MM-XFIRM at INEX Multimedia track 2007 (working notes), décembre 2007. http://inex.mmci.uni-saarland.de/static/proceedings/INEX2007-preproceedings.pdf.

[CE3] Mouna Torjmen, Karen Pinel-Sauvagnat, and Mohand Boughanem. Using pseudo-relevance feedback to improve image retrieval results. In Alessandro Nardi and Carol Peters, editors, *Advances in Multilingual and Multimodal Information Retrieval. CLEF 2007, Budapest, Hungary*, LNCS, pages 665–673. Springer, septembre 2008. https://doi.org/10.1007/978-3-540-85760-0_85.

[CE2] Lobna Hlaoua, Mouna Torjmen, Karen Pinel-Sauvagnat, and Mohand Boughanem. XFIRM at INEX 2006. Ad-hoc, Relevance Feedback and MultiMedia tracks. In Norbert Fuhr, Mounia Lalmas, and Andrew Trotman, editors, *International Workshop of the Initiative for the Evaluation of XML Retrieval (INEX), Dagstuhl, Allemagne*, volume LNCS 4518, pages 373–386,mars 2007. Springer. https://doi.org/10.1007/978-3-540-73888-6_36.

[CE1] Karen Sauvagnat, Lobna Hlaoua, and Mohand Boughanem. XFIRM at INEX 2005: adhoc and relevance feedback tracks. In Norbert Fuhr, Mounia Lalmas, Saadia Malik, and Gabriella Kazai, editors, *INitiative for the Evaluation of XML Retrieval (INEX), Dagstuhl, Germany*, volume 3977 of *LNCS*, pages 88–103, novembre 2005. Springer. https://doi.org/10.1007/978-3-540-34963-1_7.

Divers

[D2] Gilles Hubert, José Moreno, Karen Pinel-Sauvagnat, and Yoann Pitarch. Everything You Always Wanted to Know About TREC RTS* (*But Were Afraid to Ask). Diffusion scientifique, décembre 2017. https://arxiv.org/abs/1712.04671.

[D1] Linda Cappellato, Nicola Ferro, Gareth J.F. Jones, Jaap Kamps, Josiane Mothe, Karen Pinel-Sauvagnat, Eric San Juan, and Jacques Savoy. Report on CLEF 2015: Experimental IR Meets Multilinguality, Multimodality, and Interaction. *SIGIR Forum*, 49(2):(47–56, décembre 2015. http://doi.org/10.1145/2888422.2888428.

Publications nationales

Thèse

[M1] Karen Sauvagnat. *Modèle flexible pour la recherche d'information dans des corpus de documents semi-structurés*. Thèse de doctorat, Université Paul Sabatier, Toulouse, France, juin 2005. https://hal-univ-tlse3.archives-ouvertes.fr/tel-00359579, (Soutenance le 30/06/2005).

Chapitres d'ouvrages

[Ch1] Karen Pinel-Sauvagnat and Claude Chrisment. XML et recherche d'information. In Mohand Boughanem and Jacques Savoy, editors, *Recherche d'information. Etat des lieux et perspectives.*, chapter 4, pages 99–138. Hermès, http://www.editions-hermes.fr/, avril 2008. https://www.lavoisier.fr/livre/informatique/recherche-d-information-etat-des-lieux-et-perspectives/boughanem/descriptif-9782746220058, http://www.irit.fr/~Karen.Pinel-Sauvagnat/fichiers/Hermes2008.pdf, Sur invitation.

Revues nationales

[R3] Karen Pinel-Sauvagnat and Josiane Mothe. Mesures de la qualité des systèmes de recherche d'information. *Ingénierie des Systèmes d'Information, Evaluation des systèmes d'information*, Hors-série(3):11–38, 2013. http://oatao.univ-toulouse.fr/12458/.

[R2] Ines Krichen, Arlind Kopliku, Karen Pinel-Sauvagnat, and Mohand Boughanem. Une approche de recherche d'attributs pertinents pour l'agrégation d'information. *Document numérique*, 15(1):9–32, juin 2012. https://www.cairn.info/revue-document-numerique-2012-1-page-9.htm.

[R1] Karen Pinel-Sauvagnat and Mohand Boughanem. Propositions pour la pondération des termes et l'évaluation de la pertinence des éléments en recherche d'information structurée. *Information - Interaction - Intelligence*, 6(2):77–98, décembre 2006. http://www.irit.fr/~Karen.Pinel-Sauvagnat/fichiers/I32006.pdf.

Conférences nationales avec comités de sélection

[C18] Gilles Hubert, Yoann Pitarch, Karen Pinel-Sauvagnat, Ronan Tournier, and Léa Laporte. TournaRank : Quand la Recherche d'Information devient un tournoi entre documents. In *Conférence francophone en Recherche d'Information et Applications (CORIA), Rennes*, Mai 2018. Association Francophone de Recherche d'Information et Applications (ARIA).

[C17] Thomas Palmer, Gilles Hubert, and Karen Pinel-Sauvagnat. Retweeter ou ne pas retweeter: le dilemme des portails de diffusion d'information temps-réel . In *Conférence francophone en Recherche d'Information et Applications (CORIA), Marseille*, pages 123–138, mars 2017. Association Francophone de Recherche d'Information et Applications (ARIA). http://doi.org/doi:10.24348/coria.2017.28, AR= (16/33, 48%).

[C16] Rafik Abbes, Nathalie Hernandez, Karen Pinel-Sauvagnat, and Mohand Boughanem. Détection d'informations vitales pour la mise à jour de bases de connaissances . In *Journées Francophones d'Ingénierie des Connaissances (IC), Rennes*, pages 147–158, juillet 2015. Association Francaise d'Intelligence Artificielle (AFIA). https://hal.inria.fr/hal-01165507v1.

[C15] Rafik Abbes, Karen Pinel-Sauvagnat, Nathalie Hernandez, and 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), Nancy*, pages 129–143, mars 2014. LORIA. http://doi.org/doi:10.24348/sdnri.2014.CORIA-12, AR= (13/43, 30%).

[C14] Rafik Abbes, Arlind Kopliku, Karen Pinel-Sauvagnat, Nathalie Hernandez, and Mohand Boughanem. Apport du Web et du Web de Données pour la recherche d'attributs (short paper). In *Conférence francophone en Recherche d'Information et Applications (CORIA), Neuchâtel, Suisse*, pages 37–46, avril 2013. Université de Neuchâtel. http://doi.org/doi:10.24348/coria.2013.coria2013_91, AR= ((19+15)/95, 36%).

[C13]

Firas Damak, Karen Pinel-Sauvagnat, and Guillaume Cabanac. Recherche de microblogs : quels critères pour raffiner les résultats des moteurs usuels de RI ? (short paper). In *Conférence francophone en Recherche d'Information et Applications (CORIA), Bordeaux, France*, pages 317–328, mars 2012. LABRI. doi:10.24348/coria.2012.317, AR= ((14+13)/48, 56%).

[C12] Cyril Laitang, Karen Pinel-Sauvagnat, and Mohand Boughanem. Coûts de distance d'édition pour la Recherche d'Information XML . In *Conférence francophone en Recherche d'Information et Applications (CORIA), Bordeaux*, pages 357–372, mars 2012. http://doi.org/doi:10.24348/coria.2012.357, AR =(14/48, 29%).

[C11] Ines Krichen, Arlind Kopliku, Karen Pinel-Sauvagnat, and Mohand Boughanem. Une approche de recherche d'attributs pertinents pour l'agrégation d'information . In *INFormatique des Organisations et Systemes d'Information et de Decision (INFORSID), Lille*, pages 385–400, mai 2011. Association INFORSID. ftp://ftp.irit.fr/IRIT/SIG/Recherche_d_attributs_IK_AK_KPS_MB_Inforsid2011.pdf, AR= (24/82, 29%).

[C10] Arlind Kopliku, Mohand Boughanem, and Karen Pinel-Sauvagnat. Mining the Web for lists of Named Entities (short paper). In *Conférence francophone en Recherche d'Information et Applications (CORIA), Avignon*, pages 113–120, mars 2011. Association Francophone de Recherche d'Information et Applications (ARIA). http://doi.org/doi:10.24348/coria.2011.113, AR= ((16+13)/70, 41%).

[C9] Cyril Laitang and Karen Pinel-Sauvagnat. Utilisation de la théorie des graphes et de la distance d'édition pour la recherche d'information sur documents XML . In *Conférence francophone en Recherche d'Information et Applications (CORIA), Avignon*, pages 349–364, mars 2011. Association Francophone de Recherche d'Information et Applications (ARIA). http://doi.org/doi:10.24348/coria.2011.349, AR = (16/70, 23%).

[C8] Emmanuel Navarro, Yannick Chudy, Bruno Gaume, Guillaume Cabanac, and Karen Pinel-Sauvagnat. Kodex ou comment organiser les résultats d'une recherche d'information par détection de communautés sur un graphe biparti ? . In *Conférence francophone en Recherche d'Information et Applications (CORIA), Avignon*, pages 25–40, mars 2011. Association Francophone de Recherche d'Information et Applications (ARIA). http://doi.org/doi:10.24348/coria.2011.25, AR= (16/70, 23%).

[C7] Arlind Kopliku, Mohand Boughanem, and Karen Pinel-Sauvagnat. Querying by examples (poster). In *Conférence francophone en Recherche d'Information et Applications (CORIA), Sousse, Tunisie*, pages 407–408, mars 2010. Association Francophone de Recherche d'Information et Applications (ARIA). http://doi.org/doi:10.24348/coria.2010.407, AR= ((18+9+6)/70, 47%).

[C6] Mouna Torjmen and Karen Pinel-Sauvagnat. Une étude de l'impact de la structure sur la recherche multimedia . In *Conférence francophone en Recherche d'Information et Applications (CORIA), Presqu'île de Giens- Var*, pages 51–66, mai 2009. Ludovia. http://doi.org/doi:10.24348/coria.2009.51, AR= (21/60, 35%).

[C5] Mouna Torjmen, Karen Pinel-Sauvagnat, and Mohand Boughanem. Une métrique pondérée pour la recherche textuelle d'images dans des documents semi-structurés . In *Conférence francophone en Recherche d'Information et Applications (CORIA), Trégastel*, pages 55–70, mars 2008. Université de Rennes 1. http://doi.org/doi:10.24348/coria.2008.55, AR= (22/72, 30%).

[C4] Mouna Torjmen, Karen Pinel-Sauvagnat, and Mohand Boughanem. Utilisation du contexte textuel et structurel pour la recherche d'images dans des documents XML . In *Congrès Informatique des Organisations et Systèmes d'Information et de Décision (INFORSID), Perros-Guirec, 22/05/2007-25/05/2007*, pages 21–36, http://www.irisa.fr/, mai 2007. IRISA.

[C3] Lobna Hlaoua, Mohand Boughanem, and Karen Pinel-Sauvagnat. Combinaison des caractéristiques des termes pour l'extension de requêtes en recherche d'information dans les documents semi-structurés (short paper). In *Conférence francophone en Recherche d'Information et Applications (CORIA), Saint Etienne France*, pages 77–92, mars 2007. Université de Saint-Etienne. http://doi.org/doi:10.24348/coria.2007.77, AR =((21+8)/63, 45%).

[C2] Lobna Hlaoua, Mohand Boughanem, and Karen Sauvagnat. Réinjection de structures pour la reformulation de requêtes en RI structurée . In *Congrès Informatique des Organisations et Systèmes d'Information et de Décision (INFORSID), Hammamet*, pages 435–450, 2006. INFORSID (actes électroniques). http://www.irit.fr/~Karen.Pinel-Sauvagnat/fichiers/INFORSID2006.pdf.

[C1] Karen Sauvagnat and Mohand Boughanem. Propositions pour la pondération des termes et l'évaluation de la pertinence des éléments en recherche d'information structurée . In *Con-*

férence francophone en Recherche d'Information et Applications (CORIA), Lyon, pages 29–40, mars 2006. Association Francophone de Recherche d'Information et Applications (ARIA). http://doi.org/doi:10.24348/coria.2006.29, AR= (28/55, 50%).

Ateliers avec comités de sélection

[A1] Mai Dong Le and Karen Pinel-Sauvagnat. Utilisation de la distance d'édition pour l'appariement sémantique de documents XML . In *Atelier GAOC - Conférence EGC, Hammamet, Tunisie*, page (support électronique), http://www.egc.asso.fr, janvier 2010. Association Internationale Francophone d'Extraction et de Gestion des Connaissances (EGC). http://www.irit.fr/~Karen.Pinel-Sauvagnat/fichiers/GAOC2010.pdf.

Publications à caractère pédagogique

[P4] Max Chevalier, Karen Pinel-Sauvagnat, and Olivier Teste. Bases de données embarquées : intérêts et fonctionnement, exemple avec Derby. Techniques de l'ingénieur - H3867. Diffusion pédagogique, février 2014. https://www.techniques-ingenieur.fr/base-documentaire/technologies-de-l-information-th9/bases-de-donnees-42309210/bases-de-donnees-embarquees-interets-et-fonctionnement-h3867/.

[P3] Claude Chrisment, Guillaume Cabanac, Karen Pinel-Sauvagnat, Olivier Teste, and Michel Tuffery. *Bases de données orientées-objet : concepts, mise en oeuvre et exercices résolus*. Hermès Science, http://www.editions-hermes.fr/, février 2011. ISBN : 2-7462-3152-2, http://www.lavoisier.fr/livre/notice.asp?id=3LKWX3A2RA3OWG.

[P2] Max Chevalier and Karen Pinel-Sauvagnat. XML et interopérabilité des systèmes - Applications en Java 1.5 et PHP 5.0 - Techniques de l'ingénieur - H6008. Diffusion pédagogique, février 2010. https://www.techniques-ingenieur.fr/base-documentaire/technologies-de-l-information-th9/management-des-systemes-d-information-42302210/xml-et-interoperabilite-des-systemes-h6008/.

[P1] Claude Chrisment, Karen Pinel-Sauvagnat, Olivier Teste, and Michel Tuffery. *Bases de données relationnelles : concepts, mise en oeuvre & exercices*. Hermès Science, http://www.editions-hermes.fr/, juin 2008. ISBN : 978-2746220867, https://www.lavoisier.fr/livre/informatique/bases-de-donnees-relationnelles/chrisment/descriptif-9782746220867.