Contextes de lecture pondérés pour la recherche de documents structurés

Philippe Mulhem, Jean-Pierre Chevallet

UJF-Grenoble 1 / UPMF-Grenoble 2 / Grenoble INP / CNRS, LIG UMR 5217, Grenoble, F-38041, France

Philippe.Mulhem@imag.fr,Jean-Pierre.Chevallet@imag.fr

RÉSUMÉ. *Cet article s'intéresse à la recherche de parties de documents structurés, appelées doxels. Nous proposons une nouvelle notion de contexte documentaire d'un doxel que nous utilisons pour étendre un modèle de langue basé sur un lissage de Dirichlet. Nous interprétons le contexte comme une propagation du contenu des doxels sources vers le doxel en contexte dans un document. Nous montrons que cette définition de contexte est en fait une combinaison du contenu propre du doxel et de son contexte. Le modèle d'indexation proposé est compatible avec des structures à base de fichiers inverses. Nous expérimentons ce modèle sur le corpus 2009 d'INEX, en testant différents contextes. Nous mesurons une amélioration significative des résultats par rapport à une approche de référence sans contexte, pour trois types de doxels. De plus, notre proposition obtient de meilleurs résultats que le meilleur des résultats de l'évaluation Thorough de la tâche Ad Hoc d'INEX 2009.*

ABSTRACT. *This paper focuses on the retrieval of parts of structured document called doxels. We propose a notion of documentary context of a doxel and we exploit it to extend an indexing Language Model (LM) with Dirichlet smoothing. We interpret the document context of a doxel as a propagation of the content of the connected doxels via document structure links. We show that this document context definition is a combination of the intrinsic content of a doxel and its context. The new proposed indexing model is compatible with an efficient inverted files implementation. We experiment this model on INEX 2009 corpus, and test different context propagations. We measure a significant increase in results using document context, compared to a reference approach without the use of context for 3 types of doxels. Moreover, our proposal outperforms the best result obtained for the Thorough evaluation for the Ad Hoc task at INEX 2009.*

MOTS-CLÉS : *indexation, modèles de langue, lissage Dirichlet.*

KEYWORDS: *indexing, language models, Dirichlet smoothing.*

DOI:10.3166/DN.16.2.31-48

1. Introduction

Le travail présenté ici s'intéresse à la recherche de parties de documents structurés par un système de recherche d'information (SRI) orienté précision. Un *document*

structuré est un document *composé* de plusieurs parties organisées par un auteur. Cette structure est habituellement représentée par un arbre de composition. Les parties de document sont supposées être organisées de manière cohérente, car l'auteur est supposé proposer un ordre de lecture qui fournisse du sens au lecteur du document. Un système de recherche d'information qui s'attaque à des documents structurés doit proposer en réponse les parties de document les plus pertinentes à un utilisateur. Dans cet article, nous appelons *doxel* toute partie d'un document structuré qui peut être retrouvée par le SRI.

Le domaine de la recherche de documents structurés a été étudié depuis deux décennies (Wilkinson, 1994 ; Myaeng *et al.*, 1998 ; Lalmas, Ruthven, 1998). Deux grandes catégories d'approches ont été considérées dans l'état de l'art : ou bien indexer indépendamment chaque doxel et exploiter la structure des documents lors du traitement d'une requête, ou bien intégrer la prise en compte de la structure des documents lors de l'indexation.

Le travail présenté ici rentre dans la seconde catégorie d'approches. Ces approches sont selon nous, davantage en phase avec l'idée de base de la recherche d'information qui consiste à effectuer le maximum de traitements lors de l'indexation et le minimum lors du traitement des requêtes, afin de fournir les réponses le plus rapidement possible à l'utilisateur. Nous formulons la notion de *contexte de lecture* pour prendre en compte la structure d'un document par un SRI :

> Le contexte de lecture d'un doxel d est déterminé par l'ensemble de tous les doxels d' en relation avec d d'après l'auteur du document, et qui participent à l'élaboration de la signification de d par le lecteur.

Nous postulons qu'une portion de texte, considérée indépendamment du document auquel il appartient, est moins significative pour un lecteur que lorsque ce contexte est pris en compte. En fait, un contexte de lecture permet de décrire le contenu d'un doxel par celui d'autres doxels avec lesquels il est en relation. Dans ce travail, nous transposons ce postulat dans un cadre de recherche d'information (RI).

Les modèles de langues (LM) en RI sont un bon choix possible, car ils ont une modélisation théorique claire (Zhai, 2008), sont simples à implanter, et donnent de très bons résultats (Ponte, Croft, 1998). Dans le cas de la recherche de documents structurés, une utilisation simple de modèles de langues a été réalisée avec succès (Mulhem, Chevallet, 2009) lors de la campagne INEX 2009 [1] (Geva, Kamps, Trotman, 2010).

Dans les modèles de langue de recherche d'information, un document est une distribution probabiliste de termes appartenant à un langage (Zhai, 2008). Cette distribution est modélisée par la probabilité de produire la séquence de termes du document. Répondre à une requête Q consiste à retrouver le document D qui maximise la probabilité de générer Q, connaissant la distribution de la langue associée à D.

1. La tâche Ad-Hoc d'INEX 2009 propose un corpus de référence pour la recherche de documents structurés.

Il peut se révéler incorrect d'utiliser des modèles de langues pour des parties de documents, car elle peuvent être petites, mais il a été montré à INEX 2009 (Geva, Kamps, Trotman, 2010) que ces approches se comportent de manière satisfaisante. C'est pour cela que nous proposons d'utiliser de tels modèles de langue pour modéliser des doxels en contexte de lecture. La contribution de ce travail est double : d'une part elle présente une formalisation simple des contextes de lecture de doxels, ainsi que leur traduction dans des modèles de langue en définissant des propagations de pseudo-occurrences de termes ; d'autre part, un ensemble important d'expérimentations sur le corpus de référence INEX 2009 conclut à un intérêt de ces propositions.

Cet article est organisé comme suit. La section 2 présente les caractéristiques attendues des contextes de lecture. L'état de l'art est présenté en section 3. La section 4 présente notre modèle de contexte de lecture, et la manière de calculer la correspondance entre le contenu des doxels et les requêtes des utilisateurs. Des expérimentations sur INEX 2009 (Geva, Kamps, Trotman, 2010), sont présentées et commentées en partie 5, avant de conclure en section 6.

2. Contextes de lectures de doxels

Cette section se préoccupe de la définition et des caractéristiques du contexte documentaire d'un doxel. Notre objectif n'est pas ici de réaliser une présentation exhaustive du contexte dans les domaines informatiques (interaction homme-machine, langages, etc.) mais de se focaliser sur les acceptions courantes de ce terme en recherche d'information.

La recherche d'information qualifie de *contexte*, l'ensemble des éléments qui ne sont pas explicités au SRI, mais qui sont *implicites à l'interaction et à l'environnement ambiant lors de l'usage d'un SRI* (Ingwersen, Järvelin, 2005). Ce contexte peut être lié à la requête, à la tâche de l'utilisateur ou bien à des critères sociaux ou culturels. A notre avis, cette interprétation du contexte centrée sur l'utilisateur est fondamentale, mais elle ne doit pas nier l'existence d'un autre contexte, celui qui est *orienté vers les documents*, à la base du travail décrit dans cet article. Nous les appelons les *contextes de lecture*.

De manière plus précise, le contexte de lecture d'un doxel d peut être caractérisé par le fait qu'il :

1. est composé d'autres doxels, éventuellement assorti d'informations additionnelles (nature du lien, position, etc.) ;
2. doit permettre à un lecteur d'affiner son interprétation sémantique du doxel d.

Le contexte de lecture est donc composé de doxels, et nous caractérisons le contexte de lecture d'un doxel d de la manière suivante :

1. Le contexte de lecture de d est destiné à préciser sa signification : la prise en compte de ce contexte ne doit cependant pas remettre en cause l'interprétation par le système de recherche d'information du contenu initial de d, c'est-à-dire par exemple le contredire ;

2. Le contexte documentaire doit avoir un impact sur le contenu d'un doxel cible, c'est-à-dire : soit fournir une représentation additionnelle du contenu de *d*, soit modifier sa représentation initiale (*i.e.*, hors contexte documentaire) ;

3. Le contexte de lecture de *d* est sensé être spécifique à *d*, et non pas partagé par tous les autres doxels d'un document. Nous considérons en particulier que l'utilisation du corpus comme contexte documentaire de chaque doxel ne rentre pas dans le cadre des contextes documentaires tels que nous les entendons. Par exemple, la prise en compte du corpus complet pour calculer des valeurs de type *idf* est hors de notre cadre d'étude.

Dans cet article, nous nous intéressons à l'utilisation de contextes documentaires pour des documents structurés décrits avec des modèles de langue.

3. État de l'art

Les modèles de langues occupent une place de premier plan dans le domaine de la recherche d'information depuis la fin des années 1990 (Ponte, Croft, 1998). Ces modèles consistent à comparer les distributions probabilistes des documents avec celle de la requête posée par un utilisateur. Les éléments clés de ces modèles sont : 1) les variables aléatoires considérées, 2) l'hypothèse de la distribution probabiliste sous-jacente aux documents, 3) le lissage indispensable pour éviter le problème des probabilités nulles, et 4) la formule du calcul de la correspondance entre documents et requêtes. Dans le cas de textes écrits en anglais, il est couramment admis que l'utilisation d'unigrammes [2] possède des qualités indéniables : faible complexité des calculs, et très bons résultats. L'hypothèse de distribution multinomiale, proposée dans (Song, Croft, 1999), est actuellement la plus utilisée. Les lissages possibles sur les distributions de probabilités sont nombreux. Parmi eux, le lissage de Dirichlet (Zhai, Lafferty, 2001), permet de lisser les distributions de probabilités des documents en tenant compte de leur taille. Il donne de très bons résultats sans nécessiter de calculs complexes. Pour calculer la correspondance entre documents et requêtes, l'idée initiale consiste à évaluer la probabilité $P(Q|D)$, qui estime dans quelle mesure le document D peut produire la requête Q. Certaines propositions intègrent dans les modèles de langues une dépendance entre des documents. Par exemple Liu (Liu, Croft, 2004) et Shakery (Shakery, Zhai, 2008) se basent sur des documents similaires au document cible qui lui servent de contexte.

La recherche de documents structurés se confronte à deux problèmes majeurs. En premier, les parties de documents ont des tailles très variables : de quelques mots (ex. titre) à plusieurs milliers de mots (ex: article complet). En second, les parties de documents ne sont pas indépendantes les unes des autres : un auteur renforce l'argumentaire qu'il déroule au fil des phrases par une structure de texte significative. Par exemple, les ruptures typographiques (*i.e.* visuelles) formées par le groupement

2. Un unigramme est un motif composé d'un seul élément. Dans le cas des modèles de langue cela consiste à considérer les mots pris isolément, donc on ne considère aucune séquence de mots.

des phrases en paragraphes, soutiennent généralement la structuration discursive. Les titres forment une ossature terminologique, tout en guidant le lecteur par ce bornage visuel qui constitue une sémantique forte de nature « interruptive » [3]. Dans de nombreux travaux, seules sont prises en compte les dépendances hiérarchiques entre parties de documents (Pinel-Sauvagnat *et al.*, 2004 ; Myaeng *et al.*, 1998). Il est toutefois évident que des relations autres que la composition existent, même si elles sont plus subtiles et plus difficiles à contrôler. Dans une autre direction, le travail de Beigbeder (Beigbeder, 2007) propose de prendre en compte la position des termes dans les parties de documents lors du traitement de la requête. Cette approche, basée sur des ensembles flous, représente le fait que l'occurrence d'un terme dans une partie de document se propage à une autre partie de document. Cette propagation intègre une notion « d'horizon de propagation » qui dépend du contexte d'occurrence du terme. Bien que cette proposition ne se base pas dans un modèle « courant » de correspondance (comme le modèle probabiliste), elle a donné de très bons résultats lors de la campagne INEX 2008. La proposition de Arvola (Arvola *et al.*, 2011) consiste à utiliser des contextes de doxels hiérarchiques et non hiérarchiques dans un modèle de propagation de valeurs de correspondances assez complexes nécessitant de nombreuses adaptations à un corpus. De leur côté, Lv et Zhai (Lv, Zhai, 2009) ont une autre approche originale pour la prise en compte par un modèle de langue de la position des mots dans un texte, appelé PLM pour *Positional Language Model*. Ils proposent de construire un modèle de langue par position des mots dans un document. Un modèle, à une position donnée, tient compte des mots qui apparaissent dans son voisinage. Nous nous sommes inspiré de cette approche, non pas pour exprimer des distances entre termes, mais pour exprimer des distances entre doxels. Dans le cadre de modèles de langue Ogilvie et al. (Ogilvie, Callan, 2005) a proposé l'utilisation de la hiérarchie de composition, mais uniquement dans une optique de combinaison de probabilités dirigée par la hiérarchie de composition, ce qui à notre avis n'est pas suffisant.

Nous nous situons donc dans le prolongement des travaux décrits précédemment, en proposant de prendre en compte une notion plus générale de contexte qui n'est pas limitée à des dépendances hiérarchiques de structure. Notre approche est également dans la lignée de Metzler (Metzler *et al.*, 2009), sans utiliser les ancres de liens entre pages Web mais pour les documents XML : nous utilisons les contextes pour modifier les pondérations des termes d'indexation dans un doxel.

4. Modélisation du contexte de lecture des doxels

Nous étendons ici les travaux de Beigbeder (2007) et Lv (Lv, Zhai, 2009) en considérant de multiples dépendances entre doxels dans le cadre de modèles de langues. Nous ne nous intéressons pas dans cette partie à la création et aux caractéristiques de

3. Au sens où le lecteur est censé observer une pause dans sa lecture, proportionnellement à la position du titre dans la hiérarchie. Dans un roman, cette rupture a souvent un lien avec une rupture temporelle dans le récit, ou l'action du récit.

ces relations, mais uniquement à leur modélisation qui permet leur utilisation dans un SRI.

Dans la suite, nous notons DOX le corpus de doxels considéré. Pour chaque doxel $d \in DOX$ et pour chaque terme t du vocabulaire V du corpus, la fonction $C_d(t)$ dénote le nombre d'occurrences de t dans le doxel d. $C_d(t)$ est donc basée sur le contenu intrinsèque du doxel d.

4.1. Le contexte documentaire d'un doxel

Le contexte documentaire d'un doxel d, noté CDD_d, est un sous-ensemble de DOX où chaque doxel est associé à un réel strictement positif. Il s'agit donc de couples $< d', p' >$ tel que $d' \in DOX$ et $p' \in \mathbb{R}^{+*}$. Le contexte documentaire du doxel d caractérise chaque doxel d' du contexte par un poids p'. Pour faciliter les explications ultérieures, nous appellerons dans la suite d_{cont} le document d modifié par son contexte de lecture.

4.2. Les propagations de pseudo-occurrences

Nous dénotons par $V_d(t)$ le nombre de pseudo-occurrences du terme t dans un doxel d en utilisant son contexte documentaire. Ces pseudo-occurrences reflètent l'impact du contexte du doxel d. On note que $V_d(t)$ ne contient pas forcément des occurrences entières : les pseudo-occurrences sont en effet le résultat de la prise en compte des occurrences propres au doxel d, ainsi que occurrences des termes des doxels sources du contexte documentaire. Nous définissons la prise en compte de pseudo-occurrences provenant des doxels du corpus vers un doxel $d \in DOX$ par :

$$V_d(t) = C_d(t) + \sum_{<d', p'> \in CDD_d} p' \cdot C_{d'}(t) \quad (1)$$

Dans cette formule, on constate que le doxel d influence son propre vecteur de pseudo-occurrences car il est important de retrouver les occurrences de d dans ce décompte. La définition ci-dessus ne présage pas de la nature des poids p'. Il semble cependant raisonnable de leur conférer un comportement similaire à celui entre les distances entre mots des PLM de Lv (Lv, Zhai, 2009) : plus la « distance » (à définir) entre les doxels est grande, moins l'impact est important et donc plus p' est petit. Dans la suite, nous allons nous baser sur des critères simples, aisément calculables, pour évaluer ces poids.

4.3. Les modèles de langues des doxels avec contexte documentaire

Une fois le vecteur de pseudo-occurrences $V_d(t)$ défini pour tout doxel d, nous créons les modèles de doxels de la manière suivante :

$P_{ML}(t|\theta_d) = \frac{V_d(t)}{\sum_{t' \in V} V_{d(t')}}$

avec P_{ML} la probabilité pour le modèle de document θ_d de générer le terme t estimée par maximum de vraisemblance P_{ML}. Cette modélisation est cohérente avec les approches habituelles à base de modèles multinomiaux (Song, Croft, 1999).

Comme habituellement avec un modèle de langue, un lissage doit être appliqué pour éviter le problème des probabilités nulles. On peut, soit se baser sur un lissage global par le corpus DOX, soit préférer un lissage adapté à des catégories de doxels. Dans un travail précédent (Mulhem, Chevallet, 2009), une proposition d'un lissage par type de doxel a fourni de bons résultats (nous notons alors DOX_b le sous-ensemble du corpus DOX limité aux doxels de balise b). De la même manière, nous utilisons ici un lissage de type Dirichlet, avec un paramètre réel positif μ, pour modéliser le doxel d lissé par $\widetilde{\theta}_d$.

La probabilité d'un terme t dans un doxel d, lissée par le corpus DOX_b est alors calculée par :

$$P(t|\widetilde{\theta}_d) = \frac{C_d(t) + \mu \cdot P_{ML}(t|DOX_b)}{|d| + \mu} \tag{2}$$

avec $|d| = \sum_{t \in V} C_d(t)$ qui dénote la taille du doxel d; et $P_{ML}(t|DOX_b)$ l'estimation de la probabilité du terme t sur le corpus DOX_b. Dans notre cas, nous nous intéressons au modèle de langue du doxel étendu par son contexte documentaire :

$$P(t|\widetilde{\theta}_{d_{cont}}) = \frac{V_d(t) + \mu \cdot P_{ML}(t|DOX_b)}{|d_{cont}| + \mu} \tag{3}$$

avec $|d_{cont}| = \sum_{t \in V} V_d(t)$ qui dénote la taille du doxel d étendu.

4.3.1. Contexte de lecture singleton

En considérant un cas simple avec un contexte de lecture singleton, i.e. $CDD_d = \{< d', p' >\}$, nous avons :

$$\begin{aligned} P(t|\widetilde{\theta}_{d_{cont}}) &= \frac{C_d(t) + p' \cdot C_{d'}(t) + \mu \cdot P_{ML}(t|DOX_b)}{|d| + p' \cdot |d'| + \mu} \\ &= \frac{|d| + \mu}{|d| + p' \cdot |d'| + \mu} \cdot \frac{C_d(t) + \mu \cdot P_{ML}(t|DOX_b)}{|d| + \mu} \\ &\quad + \frac{p' \cdot |d'|}{|d| + p' \cdot |d'| + \mu} . \frac{p' \cdot C_{d'}(t)}{p' \cdot |d'|} \\ &= \frac{|d| + \mu}{|d| + p' \cdot |d'| + \mu} \cdot P(t|\widetilde{\theta}_d) + \frac{p' \cdot |d'|}{|d| + p' \cdot |d'| + \mu} \cdot P_{ML}(t|\theta_{d'}) \end{aligned}$$

Ce qui, en définissant une variable λ_d dénotant l'impact du contexte de lecture de d, telle que $\lambda_d = \frac{p'.|d'|}{|d|+p'.|d'|+\mu} = \frac{p'.|d'|}{|d_{cont}|+\mu}$, donne :

$$P(t|\widetilde{\theta}_{d_{cont}}) = (1 - \lambda_d) \cdot P(t|\widetilde{\theta}_d) + \lambda_d \cdot P_{ML}(t|\theta_{d'}) \tag{4}$$

4.3.2. Contexte de lecture général

Nous généralisons maintenant au cas d'un contexte de lecture non singleton pour un doxel d. En reprenant la démarche décrite ci-dessus, nous avons :

$$\begin{aligned}
P(t|\widetilde{\theta}_{d_{cont}}) &= \frac{C_d(t) + \sum_{<d',p'>\in CDD_d} p' \cdot C_{d'}(t) + \mu \cdot P_{ML}(t|DOX_b)}{|d| + \sum_{<d',p'>\in CDD_d} p' \cdot |d'| + \mu} \\
&= \frac{|d| + \mu}{|d| + \sum_{<d',p'>\in CDD_d} p' \cdot |d'| + \mu} \cdot \frac{C_d(t) + \mu \cdot P_{ML}(t|DOX_b)}{|d| + \mu} \\
&+ \frac{\sum_{<d',p'>\in CDD_d} p' \cdot |d'|}{|d| + \sum_{<d',p'>\in CDD_d} p' \cdot |d'| + \mu} \cdot \frac{\sum_{<d',p'>\in CDD_d} p' \cdot C_{d'}(t)}{\sum_{<d',p'>\in CDD_d} p' \cdot |d'|} \\
&= \frac{|d| + \mu}{|d| + \sum_{<d',p'>\in CDD_d} p' \cdot |d'| + \mu} \cdot P(t|\widetilde{\theta}_d) \\
&+ \frac{\sum_{<d',p'>\in CDD_d} p' \cdot |d'|}{|d| + \sum_{<d',p'>\in CDD_d} p' \cdot |d'| + \mu} \cdot P(t|\theta_{CDD_d}))
\end{aligned}$$

avec :

$$P(t|\theta_{CDD_d}) = \frac{\sum_{<d',p'>\in CDD_d} p' \cdot C_{d'}(t)}{\sum_{<d',p'>\in CDD_d} p' \cdot |d'|} \tag{5}$$

Dans ce cas, nous obtenons une formulation similaire au cas d'un contexte documentaire singleton :

$$P(t|\widetilde{\theta}_{d_{cont}}) = (1 - \lambda_d) \cdot P(t|\widetilde{\theta}_d) + \lambda_d \cdot P(t|\theta_{CDD_d}) \tag{6}$$

avec
$\lambda_d = \frac{\sum_{<d',p'>\in CDD_d} p' \cdot |d'|}{|d'| + \sum_{<d',p'>\in CDD_d} p' \cdot |d'| + \mu} = \frac{\sum_{<d',p'>\in CDD_d} p' \cdot |d'|}{|d_{cont}| + \mu}$

L'expression calculée en (5) mérite que l'on s'y attarde. En effet, cette expression est utile pour effectuer les calculs lors du traitement de requêtes, mais ne peut pas être considérée de manière directe comme étant une maximisation de vraisemblance sur un modèle multinomial, car les occurrences considérées des termes ne sont pas entières. Par contre, nous pouvons exprimer cette probabilité comme une combinaison des maximisations de vraisemblances des doxels d' du contexte de lecture de d par :

$$P(t|\theta_{CDD_d}) = \frac{1}{\sum_{<d',p'>\in CDD_d} p' \cdot |d'|} \cdot \sum_{<d',p'>\in CDD_d} p' \cdot |d'| \cdot \frac{p'.C_{d'}(t)}{p'.|d'|}$$

$$= \frac{1}{\sum_{<d',p'>\in CDD_d} p' \cdot |d'|} \cdot \sum_{<d',p'>\in CDD_d} p' \cdot |d'| \cdot P_{ML}(t|\theta_{d'})$$

$$= \sum_{<d',p'>\in CDD_d} \beta_{d'} \cdot P_{ML}(t|\theta_{d'})$$

Avec (en utilisant des notations p'' et d'' pour lever toute ambiguïté éventuelle) :
$\beta_{d'} = \frac{p' \cdot |d'|}{\sum_{<d'',p''>\in CDD_d} p'' \cdot |d''|}$

Cette expression confirme que nous restons dans un cadre théorique utilisant des estimations de maximisation de vraisemblance sur les doxels appartenant au contexte de lecture du doxel d.

4.4. Caractéristiques des modèles de langues

La formulation exprimée au travers de la formule (6) possède un certain nombre de caractéristiques intéressantes, au niveau théorique et au niveau pratique :

– On remarque que la prise en compte du contexte documentaire d'un élément de document structuré d peut s'interpréter de manière « contre intuitive », comme un lissage du contexte de lecture de d (estimé par maximum de vraisemblance) par combinaison linéaire avec le modèle de d lissé par le corpus ;

– Nous avons montré que la prise en compte du contexte documentaire des doxels peut être considérée comme un ensemble de termes pondérés. Une telle constatation permet une implantation efficace de ce contexte par des fichiers inverses : au lieu de représenter uniquement le poids du terme dans le doxel hors contexte documentaire, on utilise une représentation double comprenant la représentation intrinsèque et la représentation du contenu du contexte. Ensuite, lors du calcul de correspondance on utilise la représentation intrinsèque lissée et la représentation des propagations non-lissées. Il est clair que cette double représentation nécessite davantage de mémoire, mais une telle implantation a l'avantage de limiter les modifications à apporter à un SRI existant.

– Une représentation double permet également une certaine souplesse au niveau de l'expansion des contextes : si les contextes des doxels sont modifiés par ajout de nouveaux doxels, il est alors aisé de ne modifier que les champs de la représentation qui se réfèrent au contexte documentaire, sans avoir à réindexer tous les documents.

– On peut aussi choisir facilement de relativiser l'impact des contextes documentaires, en intégrant dans la formule (6) une constante α qui assigne une importance

relative a priori au contexte des doxels. Ceci a pour objectif de mieux contrôler l'impact de ces contextes, ce qui donne au final la formule suivante :

$$P(t|\widetilde{\theta}_{d_{cont}}) = (1 - \lambda_{d,\alpha}) \cdot P(t|\widetilde{\theta}_d) + \lambda_{d,\alpha} \cdot P_{ML}(t|\theta_{CDD_d}) \quad (7)$$

avec : $\lambda_{d,\alpha} = \frac{\alpha \cdot \sum_{<d',p'> \in CDD_d} p' \cdot |d'|}{|d| + \alpha \cdot \sum_{<d',p'> \in CDD_d} p' \cdot |d'| + \mu}$

La valeur α peut aussi être utilisée lors du calcul de correspondance afin de tenir compte, par exemple, de profils utilisateurs.

4.5. La correspondance entre requêtes et documents

Pour calculer la correspondance entre une requête Q et un doxel d étendu par son contexte documentaire, nous utilisons la formulation des modèles de langues qui repose sur la génération de requête par le modèle de document :

$$P(Q|\widetilde{\theta}_{d_{cont}}) = \prod_{t \in Q} P(t|\widetilde{\theta}_{d_{cont}})^{C_Q(t)} \quad (8)$$

avec $C_Q(t)$ le nombre d'occurrences du terme t dans la requête Q.

On reste donc ici dans un modèle génératif classique.

4.6. Choix de contextes des doxels

Nous nous intéressons ici à définir les paramètres importants dans le choix des contextes de doxels. Rappelons que le contexte documentaire d'un doxel est un ensemble de couples $< d', p' >$ qui indique l'impact p' d'un doxel d' sur le doxel d étendu.

Qu'attend-on du contexte documentaire d'un document ? Rappelons que nous estimons que le contexte documentaire d'un document a pour objectif de caractériser plus finement le contenu d'un document ou d'un doxel et prenant en compte d'autres documents qui lui sont liés. Ce contexte ne doit pas modifier fondamentalement le sens véhiculé par le document, ce qui se traduit en recherche d'information par une limitation de la modification de l'index. Mais comment définir le seuil au dessus duquel les modifications sont trop importantes ? L'index modifié d'un doxel peut provenir de termes ajoutés, ou bien de pondérations renforcées, et les modifications doivent donc limiter à la fois l'une et l'autre de ces caractéristiques.

Dans le cas de propagation d'index pour indexer des images dans des documents structurés, Torjmen (Torjmen *et al.*, 2010) utilise des distances basées sur les chemins dans l'arbre de composition du document. Dans notre cadre, des telles propagations ne garantissent pas la cohérence entre le doxel et son contexte documentaire, pour lequel un doxel peut apparaître structurellement proche de doxels non reliés sémantiquement.

Nous conservons cette idée, mais nous choisissons également de considérer une autre manière de définir le contexte documentaire des doxels, en choisissant pour le contexte d'un doxel d, les doxels qui lui sont déjà « proches ». Cette solution a déjà été utilisée avec succès dans (Liu, Croft, 2004) et (Shakery, Zhai, 2008), mais hors du cadre de documents structurés. Nous profitons des documents structurés pour limiter les calculs de similarités entre doxels à ceux qui appartiennent à un même document structuré. Une telle heuristique intègre donc un horizon au-delà duquel on suppose que deux doxels ne sont pas reliés. Ce choix contraint également grandement la combinatoire des calculs des liens entre les doxels d'une large collection de documents.

5. Expérimentations et résultats

Le corpus sur lequel nous effectuons nos expérimentations provient de la campagne INEX 2009 (Geva, Kamps, Trotman, 2010). Il représente actuellement la référence dans le domaine de l'indexation de documents structurés. Nous choisissons la campagne 2009 car les mesures d'évaluations proposées sont adaptées à ce que nous voulons étudier, c'est-à-dire l'impact des contextes sur la pertinence des doxels, sans considérer la taille des réponses. Le corpus d'INEX 2009 est composé de 2,5 millions de pages tirées de Wikipedia en anglais, transformées en documents XML. La mesure d'évaluation que nous utilisons est la précision interpolée au taux de rappel 0.01, notée iP[0.01]. Cette mesure est la mesure officielle de la campagne pour l'évaluation de résultats par doxels. Dans la suite, tous les résultats sont réalisés en appliquant un antidictionnaire et en transformant en minuscules toutes les lettres des doxels.

5.1. Contextes de lecture étudiés

Nous avons choisi d'étudier différents contextes pour différents types de doxels, et de fixer certains éléments pour garantir des comparaisons équitables. Nous limitons le contexte de lecture d'un doxel d aux doxels du même type qui apparaissent dans le même document que d. Nous nous intéressons à trois types de doxels : les sections (`sec`), les sous-sections de niveau 1 (`ss1`), et les sous-sections de niveau 2 (`ss2`). Ces trois types de doxels sont explicitement des doxels de structuration (alors qu'un paragraphe peut parfois être défini pour des raisons de présentation). Le nombre et la taille moyenne (en mots) sont respectivement de : 6 781 933 et 166 pour `sec`, 1 756 470 et 159 pour `ss1`, et 273 202 et 138 pour `ss2`.

Les paramètres étudiés sont de trois natures :

1. Une première direction a trait au choix des contextes. Nous avons choisi trois variantes : a) le contexte documentaire d'un doxel d contient l'ensemble des doxels de même type que d appartenant au document contenant d, noté $tout$; b) le contexte documentaire d'un doxel d contient l'ensemble des doxels de même type que d appartenant au document contenant d et apparaissant avant d dans le sens de lecture du document, noté pre; cette idée s'inspire de (Radhouani *et al.*, 2004) et est probablement bien adaptée aux documents de Wikipedia; c) le contexte documentaire d'un

doxel d contient l'ensemble des doxels de même type que d appartenant au document contenant d et apparaissant après d dans le sens de lecture du document, noté $post$; ce dernier cas n'a, à notre connaissance jamais été étudié, mais le modèle proposé permet de l'évaluer simplement.

2. La seconde évaluation porte sur le calcul des poids p' dans le contexte documentaire. Nous avons choisi deux variantes : a) l'inverse de la distance en termes de chemin minimal (l'inverse de la distance de Rada) dans le document contenant les doxels, notée par la suite *Rada* et; b) une similarité (un cosinus, notée *cos*) de contenu entre d et chaque doxel d' de son contexte. Pour comparer ces deux approches, nous avons calculé le poids moyen de ces pondérations. Pour les sections `sec`, ces moyennes sont élevées, 0,36 pour *cos* et 0,60 pour *Rada* par exemple, mais elles sont très peu corrélées : 18 % pour `ss2`, et moins de 5 % pour `sec` et `ss1`. Ces deux contextes de lectures sont donc supposés se comporter différemment.

3. Afin d'étudier l'impact des contextes documentaires, nous proposons d'utiliser quatre pondérations α décrites dans la formule (7) : ces quatre valeurs α sont 0,1 0,5 1 et 2. Pour les valeurs inférieures à 1, le doxel d a plus d'importance que son contexte pour $\alpha = 1$ l'importance est la même pour d et son contexte, pour $\alpha = 2$ le contexte de lecture est plus important.

5.2. Valeurs de lissage

Avec les modèles de langue, il a été montré l'importance du choix des valeurs de lissage (Smucker, Allan, 2006). Ce lissage a pour objectif théorique de mieux estimer les distributions de probabilités tirées des documents seuls, et au niveau pratique d'éliminer le problème des probabilités nulles dans le cas de correspondances partielles entre documents et requêtes. Dans le travail de Smucker (Smucker, Allan, 2006), les auteurs obtiennent des meilleurs résultats pour des documents textuels non-structurés pour des valeurs de lissage de 2 000. Dans nos expérimentations, nous testons les six valeurs de μ suivantes pour chaque configuration de contexte documentaire : 300, 500, 1 000, 1 500, 2 000, et 3 000.

5.3. Résultats

Les résultats présentés sont une synthèse des 450 tests effectués sur : 3 types de doxels (`ss2`, `ss2`, `sec`), 3 choix de contextes ($tout$, pre, $post$), 6 valeurs de μ (300, 500, 1 000, 1 500, 2 000, 3 000), quatre valeurs de α (0,1 0,5 1 et 2), deux pondérations de contexte documentaire (*Rada* et *cos*), en plus des 6 valeurs de μ pour les 3 références sans prise en compte de contexte informationnel. De plus, nous avons effectué des tests sans prise en compte des contextes de lecture : pour ces tests aucune pondération n'est utilisée, ces tests correspondent donc à un $\alpha = 0$ dans la formule (7). Nous évaluons grâce à ces tests plusieurs éléments :

1. la capacité de notre proposition à obtenir des meilleurs résultats en considérant les réponses par type de doxel;

2. la robustesse de notre proposition, en évaluant des tests de validation croisée sur les requêtes de la campagne INEX 2009;

3. la capacité de notre proposition à obtenir de bons résultats quand une fusion des résultats par type de doxel est effectuée, afin de fournir une seule liste en réponse.

5.3.1. Résultats par types de doxel

Nous évaluons ici l'impact positif des contextes de lecture des documents sur les résultats obtenus.

Le tableau 1 présente les meilleurs résultats obtenus par type de doxel, contexte de lecture, en faisant varier les valeurs de α et de pondération *cos* et *Rada*. Dans ce tableau, on constate que, pour chaque configuration, une amélioration importante de l'iP[0.01] se produit, entre +0,9 % et +26 %. Quand nous étudions les résultats par type de doxel, nous voyons que les meilleurs résultats sont ceux qui intègrent les doxels de tout le document (même pour les section `sec`, pour lesquelles le contexte documentaire post obtient les mêmes résultats que le contexte documentaire tout). Ce constat est particulièrement visible pour les doxels `ss2`.

Il est impossible de tirer des conclusions définitives sur les meilleures valeurs de α : les meilleurs résultats pour les `ss2` sont obtenus avec $\alpha = 1$, mais pas pour les doxels de type `sec`. On peut cependant remarquer les valeurs faibles de α ne fournissent quasiment jamais les meilleurs résultats. Il en est de même pour la valeur la plus élevée $\alpha = 2$: pour les doxels `ss1` et `ss2` cet α fournit les meilleurs résultats pour une configuration, mais jamais le meilleur résultat pour un type de doxel donné. Pour les meilleures valeurs de μ, les résultats sont également très variables : pour les types *sec* on remarque que des valeurs moyennes de μ entre 1 000 et 2 000 donnent les meilleurs résultats, mais pour les autres types ces valeurs vont de 500 à 3000.

Tous les résultats du tableau 1 sont validés en appliquant un test de significativité statistique de Student par paires. Nous évaluons l'hypothèse nulle *H0*: les couples comparés de valeurs iP[0.01], en moyenne, de la configuration sans contexte de lecture et avec contexte de lecture sont des distributions aléatoires provenant de la même distribution. Le résultat de ces tests est également présenté dans le tableau 1 : les valeurs d'iP[0.01] en gras et soulignées dénotent un rejet très significatif de *H0*, avec une valeur p<5 %, et les valeurs soulignées dénotent un rejet de *H0*, avec une valeur p<10%. Pour les doxels de type `sec` , nous notons que la majorité des contextes de lectures améliorent significativement les valeurs d'iP[0.01]. Quand la valeur d'iP[0.01] décroit pour les sections, on considère que cette différence n'est plus significative (p=40 %, largement supérieure au seuil de 5 %, malgré une augmentation de 26,4 %). Pour les doxels `ss1` une seule augmentation significative est trouvée (p=9,90 %), avec la configuration *cos* et tout le document comme contexte. Pour les doxels `ss2`, aucune augmentation significative n'est trouvée. Les résultats pour les doxels `ss2` sont bas: la raison est que nous utilisons la mesure officielle iP[0.01], basée sur la vérité terrain formées de toutes les parties de documents, et non pas uniquement les doxels de type `ss2`.

Tableau 1. Meilleurs résultats iP[0.01] par type de doxel et configuration

type de doxel	contexte	pondération	μ	α	iP[0.01] (%)
sec	aucun	aucun	1000	0	0,508 (/)
sec	*tout*	*cos*	2000	0,5	**0,579** (+14,0)
sec	*pre*	*cos*	2000	0,5	0,520 (+2,7)
sec	*post*	*cos*	2000	0,5	**0,579** (+14,0)
sec	*tout*	*Rada*	1500	1	**0,577** (+13,7)
sec	*pre*	*Rada*	1500	0,1	0,512 (+0,9)
sec	*post*	*Rada*	1500	0,5	**0,567** (+11,7)
ss1	aucun	aucun	3000	0	0,424 (/)
ss1	*tout*	*cos*	1000	1	0,486 (+14,6)
ss1	*pre*	*cos*	500	1	0,457 (+7,8)
ss1	*post*	*cos*	1000	0,5	0,465 (+9,6)
ss1	*tout*	*Rada*	1000	1	0,482 (+13,7)
ss1	*pre*	*Rada*	500	1	0,462 (+9,5)
ss1	*post*	*Rada*	1500	2	0,465 (+9,6)
ss2	aucun	aucun	500	0	0,160 (/)
ss2	*tout*	*cos*	3000	1	0,202 (+26,4)
ss2	*pre*	*cos*	3000	1	0,165 (+3,3)
ss2	*post*	*cos*	500	1	0,183 (+14,2)
ss2	*tout*	*Rada*	1500	2	0,194 (+21,2)
ss2	*pre*	*Rada*	500	0,5	0,165 (+3,4)
ss2	*post*	*Rada*	300	1	0,184 (+15,2)

5.3.2. Tests de robustesse par validation croisée

Nous avons effectué trois tests de validation croisée à deux plis pour évaluer la robustesse du choix des meilleurs paramètres de notre approche. Dans chacun de ces cas, nous avons découpé aléatoirement l'ensemble des 68 requêtes en deux sous-ensembles de 34 requêtes, appelés S_{Qa_i} et S_{Qb_i} avec $i \in 1, 2, 3$. Nous avons ensuite testé les meilleures configurations obtenues. Ces meilleures configurations groupées par type de doxel sont présentées dans le tableau 2. Dans ce dernier, les configurations en gras dénotent les meilleures configurations présentées dans le tableau 1. Pour les doxels sec, les meilleures configurations pour S_{Qa_1}, S_{Qb_1}, S_{Qa_2}, S_{Qa_3} et S_{Qb_1} diffèrent de celles du tableau 1. Cependant, ces configurations lors des tests précédents avaient obtenues des valeurs d'iP[0.01] très similaires à celles du tableau 1, amenant à la conclusion que les configurations obtenues pour les doxels sec sont stables.

Pour les doxels ss1 et ss2, les meilleures configurations sont, la plupart du temps, les mêmes de celles obtenues dans le tableau 1. Quand ce n'est pas le cas, la meilleure configuration pour la validation à deux plis correspond à la deuxième

Tableau 2. Meilleures configurations pour les valeurs d'iP[0.01] sur 3 validations croisées à deux plis

type de doxel	ensemble	configuration	iP[0.01]
sec	S_{Qa_1}	tout-Rada-1500-1	0,671
sec	S_{Qb_1}	post-Rada-1500-0.5	0,504
sec	S_{Qa_2}	tout-Rada-1500-1	0,600
sec	S_{Qb_2}	**tout-cos-2000-0,5**	0,573
sec	S_{Qa_3}	tout-Rada-1500-1	0,528
sec	S_{Qb_3}	tout-Rada-1500-1	0,607
ss1	S_{Qa_1}	**tout-cos-1000-1**	0,520
ss1	S_{Qb_1}	tout-Rada-1000-1	0,460
ss1	S_{Qa_2}	**tout-cos-1000-1**	0,516
ss1	S_{Qb_2}	tout-Rada-1000-1	0,461
ss1	S_{Qa_3}	**tout-cos-1000-1**	0,437
ss1	S_{Qb_3}	**tout-cos-1000-1**	0,482
ss2	S_{Qa_1}	**tout-cos-3000-1**	0,239
ss2	S_{Qb_1}	post-Rada-300-1	0,175
ss2	S_{Qa_2}	tout-Rada-300-1	0,205
ss2	S_{Qb_2}	**tout-cos-3000-1**	0,207
ss2	S_{Qa_3}	**tout-cos-3000-1**	0,189
ss2	S_{Qb_3}	**tout-cos-3000-1**	0,164

meilleure (pour les doxels ss1), ou la troisième meilleure pour les doxels ss2, toutes très proches des meilleures configurations.

Nous concluons donc que notre approche est stable.

5.3.3. Intégration des résultats des doxels de types sec*,* ss1 *et* ss2*.*

Nous utilisons ici une approche qui a prouvé son efficacité pour la recherche de documents structurés, l'approche $Fetch$ and $Browse$. La stratégie utilisée, décrite par (Lalmas, 1997), est décomposée en deux étapes : 1) pendant le $Fetch$ les documents entiers sont comparés à la requête et triés selon leur valeur de pertinence, puis 2) l'étape de $Browse$ prend les documents de la liste du $Fetch$ pour trouver leurs parties les plus pertinentes.

Une approche basée sur cette idée a obtenu les meilleurs résultats de l'évaluation $Ad\ Hoc\ Thorough$ de la campagne INEX 2009 (Geva, Kamps, Lethonen *et al.*, 2010). Nous avons réalisé une étape $Fetch$ en utilisant un modèle langue utilisant un lissage de Dirichlet (paramètre de lissage μ=900). L'étape de $Browse$ consiste à rechercher les meilleurs parties de documents, en recherchant les doxels de type sec,

ss1 et ss2. Quand nous utilisons les meilleures configurations du tableau 1 pour les doxels de type sec, ss1 and ss2, à la suite de l'étape de $Fetch$, le résultat que obtenons est une valeur d'iP[0.01] de 0,6741. Cette valeur est 6,4 % plus élevée que les meilleurs résultats officiels d'INEX 2009, avec des valeurs d'iP[0.01] de 0,6333 par l'Université de Waterloo (deux premiers résultats) est plus élevée de 9,8 % que le troisième résultat, iP[0.01]=0,6141, par l'Université Pierre et Marie Curie (Geva, Kamps, Lethonen *et al.*, 2010). Sans les résultats officiels fournis lors de cette campagne, il est impossible d'effectuer des tests de significativité statistique. Cependant, il est courant en recherche d'information de considérer qu'une amélioration de 5 % dénote une différence significative. Même sans pouvoir conclure sur la significativité des différences, nous constatons que notre proposition fournit des résultats meilleurs pour la recherche de documents structurés.

6. Conclusion

Nous avons défini dans cet article la notion de contexte de lecture de doxels. Notre modélisation de ces contextes permet de les utiliser simplement dans des modèles de langue de recherche d'information, en respectant les fondamentaux de ces modèles. Nous avons testé de nombreuses variations de cette proposition, avec différents schémas de pondération (basés sur la distance dans l'arbre de la structure ou sur le contenu) sur le corpus INEX 2009. Les résultats obtenus sur les type de doxels sec, ss1 et ss2 montrent que l'utilisation de contexte de lecture améliore significativement les résultats. Les contextes de lectures proposés, limités aux doxels qui appartiennent au même document, génèrent un coût en $O(m^2/2)$ avec m nombre moyen de doxels d'un type par document. Ces expérimentations montrent que :

– prendre en compte tous les doxels du même type dans un document fournit de meilleurs résultats que le fait de ne pas utiliser de contexte ;

– prendre davantage en compte le contexte de lecture pour les petits doxels donne de meilleurs résultats, même si les différences de qualité que nous avons obtenues ne sont pas significatives statistiquement ;

– intégrer les contextes documentaires à la suite d'une opération de $Fetch$ donne de très bons résultats, meilleurs que les meilleurs résultats officiels d'INEX 2009. Ceci prouve que des approches $Fetch$ et $Browse$ peuvent être améliorées par l'utilisation des contextes de lecture proposés.

L'étude des relations entre doxels que nous avons faite ne propose pas de prendre en compte de manière conjointe plusieurs types de relations entre doxels. Une extension simple pourrait alors estimer si de telles combinaisons sont utiles pour obtenir de meilleurs résultats. Les pondérations utilisées sont directement prises en compte dans le calcul des distributions des termes dans les doxels en contexte de lecture, on peut se poser la question de savoir si des modélisations plus fines pourraient également donner de meilleurs résultats. Le travail présenté ici s'est limité à des doxels reliés directement. Nous pouvons étendre ce cas à des doxels connectés transitivement, en

se rapprochant alors de travaux similaires à ceux de Pagerank, mais cette fois utilisés lors de l'indexation.

Bibliographie

Arvola P., Kekalainen J., Junkkari M. (2011). Contextualization models for xml retrieval. *Information Processing & Management*, vol. 47, n° 5, p. 762 - 776.

Beigbeder M. (2007). Structured content-only information retrieval using term proximity and propagation of title terms. In N. Fuhr, M. Lalmas, A. Trotman (Eds.), *Comparative evaluation of xml information retrieval systems*, vol. 4518, p. 200-212. Springer Berlin / Heidelberg.

Geva S., Kamps J., Lethonen M., Schenkel R., Thom J., Trotman A. (2010). Overview of the inex 2009 ad hoc track. In S. Geva, J. Kamps, A. Trotman (Eds.), *Focused retrieval and evaluation*, vol. 6203, p. 4-25. Springer Berlin / Heidelberg.

Geva S., Kamps J., Trotman A. (2010). Focused retrieval and evaluation, 8th international workshop of the initiative for the evaluation of xml retrieval, inex 2009, brisbane, australia, december 7-9, 2009, revised and selected papers. In *Inex*, vol. 6203. Springer.

Ingwersen P., Järvelin K. (2005, December). Information retrieval in context: Irix. *SIGIR Forum*, vol. 39, p. 31–39.

Lalmas M. (1997). Dempster-shafer's theory of evidence applied to structured documents: Modelling uncertainty. In *Sigir*, p. 110-118.

Lalmas M., Ruthven I. (1998). Representing and retrieving structured documents using the dempster-shafer theory of evidence: Modelling and evaluation. *JOURNAL OF DOCUMENTATION*, vol. 54, p. 529–565.

Liu X., Croft W. B. (2004). Cluster-based retrieval using language models. In *Proceedings of the 27th annual international acm sigir conference on research and development in information retrieval*, p. 186–193. New York, NY, USA, ACM. Consulté sur http://doi.acm.org/10.1145/1008992.1009026

Lv Y., Zhai C. (2009). Positional language models for information retrieval. In *Proceedings of the 32nd international acm sigir conference on research and development in information retrieval*, p. 299–306. New York, NY, USA, ACM. Consulté sur http://doi.acm.org/10.1145/1571941.1571994

Metzler D., Novak J., Cui H., Reddy S. (2009). Building enriched document representations using aggregated anchor text. In *Proceedings of the 32nd international acm sigir conference on research and development in information retrieval*, p. 219–226. New York, NY, USA, ACM.

Mulhem P., Chevallet J.-P. (2009). Use of language model, phrases and wikipedia forward links for inex 2009. In *Inex*, p. 103-111.

Myaeng S. H., Jang D.-H., Kim M.-S., Zhoo Z.-C. (1998). A flexible model for retrieval of sgml documents. In *Proceedings of the 21st annual international acm sigir conference on research and development in information retrieval*, p. 138–145. New York, NY, USA, ACM. Consulté sur http://doi.acm.org/10.1145/290941.290980

Ogilvie P., Callan J. (2005). Hierarchical language models for xml component retrieval. In N. Fuhr, M. Lalmas, S. Malik, Z. Szlávik (Eds.), *Advances in xml information retrieval*, vol. 3493, p. 269-285. Springer Berlin / Heidelberg.

Pinel-Sauvagnat K., Boughanem M., Chrisment C. (2004, octobre). Searching XML documents using relevance propagation (regular paper). In A. Apostolico, M. Melucci (Eds.), *Symposium on String Processing and Information Retrieval (SPIRE), Padoue, Italie, 06/10/2004-08/10/2004*, p. 242–254. http://www.springerlink.com, Springer. Consulté sur http://www.irit.fr/~Karen.Pinel-Sauvagnat/fichiers/XFIRM@SPIRE04.pdf

Ponte J. M., Croft W. B. (1998). A language modeling approach to information retrieval. In *Proceedings of the 21st annual international acm sigir conference on research and development in information retrieval*, p. 275–281. New York, NY, USA, ACM. Consulté sur http://doi.acm.org/10.1145/290941.291008

Radhouani S., Chevallet J.-P., Géry M. (2004). Un modèle à base de chemin de lecture pour la recherche d'informations précises sur le web. In *Coria*, p. 249-270.

Shakery A., Zhai C. (2008, April). Smoothing document language models with probabilistic term count propagation. *Inf. Retr.*, vol. 11, p. 139–164. Consulté sur http://portal.acm.org/citation.cfm?id=1349658.1349664

Smucker M. D., Allan J. (2006, November). *An investigation of dirichlet prior smoothing's performance advantage*. Rapport technique n° CIIR Technical Report IR-548. University of Massachusetts.

Song F., Croft W. B. (1999). A general language model for information retrieval. In *Proceedings of the eighth international conference on information and knowledge management*, p. 316–321. New York, NY, USA, ACM. Consulté sur http://doi.acm.org/10.1145/319950.320022

Torjmen M., Pinel-Sauvagnat K., Boughanem M. (2010, octobre). 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*, vol. 5, n° 4, p. 323–352. Consulté sur http://dx.doi.org/10.1504/IJBIDM.2010.036123 ,http://www.irit.fr/~Karen.Pinel-Sauvagnat/fichiers/IJBIDM2009.pdf

Wilkinson R. (1994). Effective retrieval of structured documents. In *Proceedings of the 17th annual international acm sigir conference on research and development in information retrieval*, p. 311–317. New York, NY, USA, ACM. Consulté sur http://dl.acm.org/citation.cfm?id=188490.188591

Zhai C. (2008). *Statistical language models for information retrieval*. Hanover, MA, USA, Now Publishers Inc.

Zhai C., Lafferty J. (2001). A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proceedings of the 24th annual international acm sigir conference on research and development in information retrieval*, p. 334–342. New York, NY, USA, ACM. Consulté sur http://doi.acm.org/10.1145/383952.384019