Un modèle de contexte documentaire par doxels pondérés.

Application à un modèle de langue contextuel 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,Jean-Pierre.Chevallet}@imag.fr

RÉSUMÉ. *Cet article porte sur la recherche de parties de documents appelées doxels. Nous définissons la notion de contexte documentaire d'un doxel, en utilisant deux éléments : 1) un lissage de type Dirichlet sur des doxels, et 2) une interprétation du contexte d'un doxel par des propagations du contenu des autres doxels de contexte. Nous montrons que cette interprétation de contexte documentaire est exprimable par des combinaisons du contenu intrinsèque lissé et des contenus propagés, non lissés, des doxels. Cette expression est donc compatible avec des implantations à base de fichiers inverses. Nous avons étudié différentes propagations sur le corpus INEX 2009, en constatant systématiquement une forte augmentation des résultats en contexte, par rapport à une approche par modèle de langue de référence hors contexte.*

ABSTRACT. *This article focuses on the retrieval of documents parts called doxels. We define the notion of documentary context of a doxel by exploring two elements: 1) a Dirichlet smoothing on doxels, and 2) interpretating the documentary context of a doxel by propagating the content of the context doxels. We show that this interpretation of documentary context can be expressed by a combination of the intrinsic content and the content, unsmoothed, of the context doxels. This expression is therefore compatible with inverted files implementation. We studied several propagations on the INEX corpus 2009, and we found a significant increase in systematic results using the documentary contexts, compared to a reference approach without context.*

MOTS-CLÉS : *Indexation, Recherche, XML, INEX.*

KEYWORDS: *Indexing, Retrieval, XML, INEX.*

CORIA 2012, pp. 341–356, Bordeaux, 21-23 mars 2012

1. Introduction

Le travail présenté ici se positionne dans le contexte de l'accès à des parties de documents structurés par un Système de Recherche d'Information (SRI) orienté vers la précision des réponses. Un document structuré est un document *composé* de différentes parties, organisées par leur auteur. Cette organisation prend habituellement la forme d'un arbre de composition qui peut être exprimé dans le langage XML (eXtensible Modeling Language). Nous désignons par *doxel* tout élément composant un document et susceptible d'être retrouvé par le SRI. Ces parties de documents sont supposées cohérentes entre elles car leur structure a été définie par l'auteur et leur ordonnancement est sensé suivre un chemin de lecture qui doit faire sens pour le lecteur.

Afin de retrouver les parties de documents (i.e. doxels) les plus pertinentes, plusieurs catégories d'approches sont possibles. On peut indexer les doxels indépendamment les uns des autres, et tenir compte de la structure des documents qu'au moment du traitement des requêtes. Au contraire, le SRI peut exploiter les relations entre les doxels lors de l'indexation. Les travaux décrits dans cet article, se basent sur la prise en compte de la structure des documents lors de l'indexation. Nous considérons que cette approche est plus en cohérence avec les SRI, pour lesquels le maximum de traitements doit être réalisés lors de l'indexation afin de réduire au maximum les calculs lors de la phase d'interrogation pour ainsi fournir des réponses le plus rapidement possible à l'utilisateur. La prise en compte de la structure des documents est la base d'une notion de *contexte documentaire* de doxels que nous proposons dans cet article. Nous définissons un contexte documentaire d'un doxel comme l'ensemble des parties d'un document qui lui sont reliés et qui participent pour le lecteur, à l'élaboration de sa signification. Nous partons du postula qu'un texte extrait du contexte de l'ouvrage auquel il appartient, a moins de signification que lorsque qu'il reste plongé dans son contexte documentaire. Le contexte documentaire a ainsi pour objectif de compléter le contenu d'un doxel et prenant en compte d'autres éléments qui lui sont liés.

Une fois la représentation d'un contexte documentaire défini, il faut s'intéresser au modèle du recherche d'information à mettre en oeuvre pour le calcul de la correspondance. Dans le domaine de la recherche d'information, les approches à base de modèles de langues sont très utilisés dans les recherches actuelles, en raison d'une modélisation claire, de leur simplicité de mise en oeuvre et bien sûr, en raison de la qualité des résultats obtenus (Ponte *et al.,* 1998). Dans le cadre des documents structurés, une utilisation simple de ces modèles de langue à déjà été proposée et mise en oeuvre (Mulhem *et al.,* 2009) lors de la campagne d'évaluation INEX 2009 (Geva *et al.,* 2010). Il peut paraître étrange d'appliquer ces modèles de langues pour une indexation de documents structurés car ces modèles construisent les index des documents en estimant la distribution statistique des termes à partir du contenu des documents. Ce modèle statistique est alors une sorte de signature de la langue utilisée par l'auteur pour s'exprimer dans ce document. Tout document qui semble "produire" la requête par ce modèle de langue, est alors déclaré pertinent. On se doute alors que considérer des doxels revient à considérer des "documents" de très petite taille, comme des titres avec quelques dizaines de mots. Il paraît alors abusif d'exploiter un modèle de

langue dans ce cas. Pourtant, les bons résultats que nous avons eu précédemment nous ont conforté dans cette voie de recherche, et nous avons alors décidé d'exploiter le contexte de ces doxels, d'une part pour contrecarrer leur faible taille, mais aussi parce que nous pensons que le contexte participe à la construction de leur signification par le lecteur. Ainsi, nous proposons dans cet article, d'explorer plus en avant l'usage d'un modèle de langue pour l'indexation de doxels en exploitant un contexte documentaire.

Le plan de cet article est le suivant. La section 2 présente une description des caractéristiques attendues des contextes documentaires. Un état de l'art sur les travaux connexes à notre étude est présenté en section 3. Nous décrivons en section 4 notre proposition de modèle de documents structurés, en reformulant les calculs pour obtenir une expression facilitant l'utilisation des contextes des doxels par un système de recherche d'information. Une discussion sur la manière de déterminer le contexte documentaire de doxels dans des documents structurés est décrite en section 5. Le modèle défini a été implanté et les résultats sur le corpus de la campagne d'évaluation INEX 2009 (Geva *et al.,* 2010), référence dans le domaine, sont présentés et discutés en partie 6, avant de conclure en partie 7.

2. Le contexte documentaire d'un doxel

Cette partie, 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 d'usage d'un SRI (Ingwersen *et al.,* 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 centrée utilisateur du contexte 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 documentaires*.

De manière plus précise, le contexte documentaire d'un doxel d :

1) est composé d'autres doxels, éventuellement assorti d'informations additionnelles (nature du lien, position, etc) ;

2) doit permettre d'affiner l'interprétation sémantique de d par le lecteur.

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

– Le contexte documentaire 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 ;

– Le contexte documentaire doit impacter 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).

– Le contexte documentaire 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 la 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.

Ainsi, dans cet article, nous nous intéressons à l'utilisation de contextes documentaires pour des documents structurés décrits avec des modèles de langue. La section suivante dresse un rapide état de l'art sur ces domaines.

3. Etat 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 90 (Ponte *et al.,* 1998). Ces modèles consistent à comparer les distributions probabilistes des documents et celles 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[1] possède des qualités indéniables : faible complexité des calculs, et très bons résultats. L'hypothèse de distribution multinomiale, proposée dans (Song *et al.,* 1999), est actuellement la plus utilisée. Les lissages possibles sur les distributions de probabilités sont nombreux. Parmi eux, le lissage de Dirichlet (Zhai *et al.,* 2001), permet de lisser les 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 *et al.,* 2004) et Shakery (Shakery *et al.,* 2008) se basent sur des documents similaires au document cible qui lui servent de contexte. L'approche proposée dans cet article suit cette idée, en l'appliquant aux documents structurés.

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

1. 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.

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 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"[2]. 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 intégre 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), il a donné de très bons résultats lors de la campagne INEX 2008. Les auteurs de (Arvola *et al.,* 2011) ont également proposé d'utiliser des contextes de doxels hiérarchiques et non-hiérarchiques dans un modèle de propagation de valeurs de correspondances assez complexe, dépendant des contextes, qui est difficile à étudier. De leur coté, Lv et Zhai (Lv *et al.,* 2009) ont une autre approche originale pour la prise en compte par un modèle de langue de la position des mots dans un texte. Ils proposent carrément 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.,* 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 n'est pas à notre avis suffisant. La partie suivante détaille notre proposition de modèle.

4. Un modèle de Contexte Documentaire par Doxels Pondérés (CDDP)

Le modèle que nous proposons se situe dans la mouvance des travaux de Beigbeder et Lv (Beigbeder, 2007, Lv *et al.,* 2009) : nous voulons prendre en compte de multiples dépendances entre les parties de documents (doxels), dans un contexte de modèle de langue. Nous ne nous intéressons pas à la caractérisation ni à l'extraction des relations dans un document structuré, mais seulement à leur modélisation pour permettre leur utilisation par un système de recherche d'information.

---

2. au sens où le lecteur est sensé 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.

Dans la suite, nous dénotons par DOX le corpus de doxels considéré. Pour chaque doxel $d \in DOX$ et pour chaque terme t du vocabulaire V du corpus, un vecteur d'occurrence initial $C_d(t)$ dénote le nombre d'occurrences de t dans le doxel d. Cette valeur $C_d(t)$ 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 ensemble de couples $< d, p >$ tel que $d \in DOX$ et $p \in \mathbb{R}$ (réel). 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 documentaire.

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 le vecteur $V_{d'}(t)$ ne contient pas forcément des occurrences entières : les pseudo-occurrences sont en effet le résultat de la prise en compte du vecteur d'occurrence initial du doxel, ainsi que les vecteurs initiaux 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 :

$$V_{d'}(t) = C_{d'}(t) + \sum_{<d,p> \in CDD_{d'}} p.C_d(t) \qquad [1]$$

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 *et al.,* 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 $\theta_{d'}$ de générer le terme t estimée par maximum de vraisemblance P_{ML}.

Cette formulation, courante dans les modèles de langues, n'implique aucune hypothèse. Elle est cependant cohérente avec les approches habituelles à base de modèles multinomiaux (Song *et al.,* 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, soit préférer un lissage adapté à des catégories de doxels. Dans un travail précédent (Mulhem *et al.,* 2009), une proposition d'un lissage par type de doxel a fourni de bons résultats. Pour cette raison, nous utilisons ici un lissage de type Dirichlet 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 est alors calculée par :

$$P(t|\widetilde{\theta}_{d'}) = \frac{C_{d'}(t) + \mu.P_{ML}(t|DOX)}{|d'| + \mu} \quad [2]$$

avec $|d'| = \sum_{t \in V} C_{d'}(t)$ qui dénote la taille du doxel d'. Dans notre cas, nous nous intéressons au modèle de langue du doxel étendu par son contexte documentaire :

$$P(t|\widetilde{\theta}_{d'_{cont}}) = \frac{V_{d'}(t) + \mu.P_{ML}(t|DOX)}{|d'_{cont}| + \mu} \quad [3]$$

avec $|d'_{cont}| = \sum_{t \in V} V_{d'}(t)$ qui dénote la taille du doxel d' étendu.

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

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

Ce qui, en définissant une variable $\lambda_{d'}$ dénotant l'impact du contexte documentaire de d', telle que $\lambda_{d'} = \frac{p.|d|}{|d'|+p.|d|+\mu} = \frac{p.|d|}{|d'_{cont}|+\mu}$:

$$P(t|\widetilde{\theta}_{d'_{cont}}) = (1 - \lambda_{d'}).P(t|\widetilde{\theta}_{d'}) + \lambda_{d'}.P_{ML}(t|\theta_d) \quad [4]$$

Une généralisation [3] à un contexte documentaire non singleton pour un doxel d', donne :

$$P(t|\widetilde{\theta}_{d'_{cont}}) = (1 - \lambda_{d'}).P(t|\widetilde{\theta}_{d'}) + \lambda_{d'}.P_{ML}(t|\theta_{CDD_{d'}}) \quad [5]$$

avec : $P_{ML}(t|\theta_{CDD_{d'}}) = \frac{\sum_{<d,p>\in CDD_{d'}} p.C_d(t)}{\sum_{<d,p>\in CDD_{d'}} p.|d|}$, et
$\lambda_{d'} = \frac{\sum_{<d,p>\in CDD_{d'}} p.|d|}{|d'|+\sum_{<d,p>\in CDD_{d'}} p.|d|+\mu} = \frac{\sum_{<d,p>\in CDD_{d'}} p.|d|}{|d'_{cont}|+\mu}$.

4.4. *Caractéristiques des modèles de langues*

La formulation exprimée par formule [5] possède un certain nombre de caractéristiques intéressantes, aussi bien à un niveau théorique qu'à un 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 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 Système de Recherche d'Information (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 [5] 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 :

$$P(t|\widetilde{\theta}_{d'_{cont}}) = (1 - \lambda_{d',\alpha}).P(t|\widetilde{\theta}_{d'}) + \lambda_{d',\alpha}.P_{ML}(t|\theta_{CDD_{d'}}) \quad [6]$$

avec : $\lambda_{d',\alpha} = \frac{\alpha.\sum_{<d,w>\in CDD_{d'}} p.|d|}{|d'|+\alpha.\sum_{<d,w>\in CDD_{d'}} p.|d|+\mu}$.
On peut noter que la valeur α peut aussi être utilisée lors du calcul de correspondance afin de tenir compte, par exemple, de profils utilisateurs.

---

3. Non décrite ici par manque de place.

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 :

$$P(Q|\widetilde{\theta}_{d'_{cont}}) = \prod_{t \in Q} P(t|\widetilde{\theta}_{d'_{cont}})^{C_Q(t)} \quad [7]$$

5. Discussion et choix sur la définition des doxels en contexte

Nous nous intéressons ici à définir quelques paramètres qui semblent 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 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 un 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 *et al.,* 2004) et (Shakery *et al.,* 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. Un tel choix intègre donc à la similarité, l'hypothèse que les doxels d'un même documents portent sur les mêmes sujets ou des sujets connexes. Ce choix limite également grandement la combinatoire quand au calcul des liens entre tous les doxels d'une large collection de documents.

6. Experimentations et résultats

Le corpus sur lequel nous effectuons nos expérimentations provient de la campagne INEX 2009 (Geva *et al.,* 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 documents pertinents 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 moyenne (MAiP) qui est une MAP modifiée pour s'appliquer aux recherches de documents structurés. Dans la suite, tous les résultats sont réalisés en appliquant un antidictionnaire et en transformant en minuscules toutes les lettres des doxels.

6.1. *Contextes étudiés*

Nous avons choisi d'étudier différents contextes pour différents types de doxels, et de fixer certains éléments pour garantir des comparaisons équitables. L'élément majeur fixé pour les expérimentations est de limiter les contextes documentaires à ceux appartenant à un même document. Cette limitation va nous permettre de mesurer l'apport de contextes simples. Nous choisissons d'effectuer nos tests sur 3 types de doxels : les sections (sec), sous-sections de niveau 1 (ss1) et 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). Ce choix permet en outre de vérifier dans quelle mesure notre proposition s'adapte à des doxels de tailles différentes. Les paramètres étudiés sont de deux 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' ; 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 ; 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 ; 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 la calcul du poids p dans le contexte documentaire. Nous avons choisi deux variantes : a) l'inverse de la distance en terme de chemin minimal (1/distance de Rada) dans le document contenant les doxels ; b) une similarité (un cosinus) de contenu entre d' et chaque doxel de son contexte.

Afin d'étudier l'impact des contextes documentaires, nous proposons d'utiliser deux pondérations α décrites dans la formule [4] de la partie 4.3. : ces deux valeurs sont 1 et 0,5 . Dans le premier cas on utilise les nombres d'occurrences des documents

du contexte tels quels, dans le second on divise par deux ces nombres d'occurrences pour accorder moins d'importance au contexte.

6.2. *Valeurs de lissage*

Avec les modèles de langue, il a été montré l'importance du choix des valeurs de lissage (Smucker *et al.,* 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 (Smucker *et al.,* 2006), les auteurs obtiennent des meilleurs résultats pour des documents textuels non-structurés pour des valeurs de lissage de 2000. Dans nos expérimentations, nous testons les six valeurs de μ suivantes pour chaque configuration de contexte documentaire : 300, 500, 1000, 1500, 2000, et 3000.

6.3. *Résultats*

Les résultats présentés sont une synthèse des 72 évaluations (3 choix de contextes (document, document-avant, document-après), 6 valeurs de μ (300, 500, 1000, 1500, 2000, 3000), deux valeurs de α (0,5 et 1), deux pondérations de contexte documentaire (Rada, et cosinus)) menées pour chacun des 3 types de doxels considérés (ss2, ss1, sec), en plus des 6 valeurs de μ pour les 3 références sans prise en compte de contexte informationnel, donnant un total de 234 expérimentations.

6.3.1. *Meilleurs résultats par configuration*

Dans cette partie, nous nous intéressons aux meilleurs résultats obtenus par type de doxel : nous cherchons à savoir si l'utilisation des contextes documentaires tels que nous les avons définis dans cet article, donnent de meilleurs résultats qu'une approche courante qui considère chaque doxels indépendamment.

Nous constatons dans le tableau 1 que, pour tous les types de doxels considérés et quelque soit le contexte documentaire utilisé, la progression en terme de MAiP (indiquées entre parenthèses dans dernière colonne) est très importante, entre 10 et 74%. En regardant plus en détail ces résultats, les contextes globaux, c'est-à-dire ceux utilisant tous les doxels de même type dans un document, donnent de meilleurs résultats, que ceux qu'utilisent uniquement les doxels avant ou bien les doxels après le doxel en contexte (sauf pour les contextes des doxels de type *sec* où les meilleurs résultats sont les mêmes). Ceci est particulièrement notable pour les doxels de petite taille, telle que ceux de type *ss2*. La valeur de α qui fournit le meilleur résultat est le plus souvent 1.0. Les valeurs de μ donnant les meilleurs résultats sont souvent comprises entre 1500 et 2000, sauf pour les contextes documentaires de ss2.

Ce résultat montre clairement que notre modèle à base de contextes documentaires est un bon choix pour améliorer les résultats d'une indexation de doxels dans le cadre

Tableau 1. *Meilleurs résultats par configuration de contexte documentaire par type de doxel*

type de doxel	doxel de contexte	contexte	μ	α	MAiP (%)
sec	/	/	1000	/	0.1342 (/)
sec	tous	cosinus	1500	0,5	0.1908 (+42.16%)
sec	avant	cosinus	2000	0,5	0,1537 (+14,52%)
sec	après	cosinus	1500	0,5	0,1908 (+42,16%)
sec	tous	Rada	1500	0,5	0,1952 (+45,40%)
sec	avant	Rada	1000	1.0	0,1489 (+10,94%)
sec	après	Rada	1500	0.5	0,1952 (+45,40%)
ss1	/	/	1500	/	0,056 (/)
ss1	tous	cosinus	2000	1,0	0,097 (+73,67%)
ss1	avant	cosinus	1000	0,5	0,079 (+42,23%)
ss1	après	cosinus	1500	0,5	0,094 (+68,27%)
ss1	tous	Rada	1500	1,0	0,096 (+72,35%)
ss1	avant	Rada	1000	1,0	0,079 (+42,87%)
ss1	après	Rada	1500	0,5	0,091 (+63,51%)
ss2	/	/	500	/	0,0106 (/)
ss2	tous	cosinus	1500	1.0	0,0178 (+66,5%)
ss2	avant	cosinus	300	1.0	0,0142 (+32,68%)
ss2	après	cosinus	300	1.0	0,0141 (+32,01%)
ss2	tous	Rada	300	1.0	0,0164 (+53,56%)
ss2	avant	Rada	500	0.5	0,0138 (+29,20%)
ss2	après	Rada	300	1.0	0,0138 (+29,47%)

d'une collection de documents structurés. Ces résultats montrent également qu'il vaut mieux considérer le contexte du sens de lecture en aval (les doxels qui suivent dans le texte le doxel d), qu'en amont. Ce résultat nous surprend, car il nous semble contre intuitif. En effet, le contenu de ce qui précède un doxel d, nous semble plus important pour comprendre le sens de ce doxel, car justement ce qui le précède est sensé être lu *avant*. Nous n'avons actuellement pas de piste pour expliquer ce résultat.

6.3.2. *Étude de la configuration (ss1, document, cosinus)*

Nous choisissons maintenant de nous focaliser sur l'expansion sur les sous-sections de niveau 1, en examinant le comportement des contextes de tout le document. La dernière colonne du tableau 2 présente (entre parenthèses) l'augmentation relative des résultats par rapport à une indexation sans le contexte documentaire, égale à 0,056 (cf. le tableau 1). Dans ce cas, on remarque que l'utilisation des contextes avec un α égal à 1 est de manière constante, supérieure à l'utilisation des $\alpha = 0,5$. Dans ce cas, les résultats sont meilleurs quand l'impact des contextes est plus important, et donc quand on utilise davantage d'informations pour représenter les sous-sections. Si

l'on s'intéresse aux variations des résultats pour chaque valeur de α en fonction des valeurs de lissage μ, le comportement est similaire : une augmentation de la MAiP de 300 à 1500, une stagnation entre 1500 et 2000, puis une chute entre 2000 et 3000. On constate donc ici des comportements similaires à ce qui est rapporté dans (Smucker *et al.,* 2006) pour des textes non structurés. Sur ces corpus, ce constat est vérifié sauf pour les doxels très petits comme ceux de type ss2. Un autre comportement notable est apparu dans beaucoup de nos expérimentations : pour des configurations de contextes documentaires identiques, les valeurs de lissages les meilleures pour un α le sont aussi pour l'autre. Bien entendu, nos expérimentations n'ayant portées que sur deux valeurs, cela ne peux pas être considéré comme un comportement significatif. Cependant, si un tel comportement se confirme, et si les valeurs des α peuvent varier entre différents utilisateurs (cf. section 4.4), un tel résultat serait très intéressant car nous n'aurions pas à modifier les lissages entre les utilisateurs.

Tableau 2. *Résultats pour contextes ss1 avec relations pondérées par cosinus*

μ	α	MAiP (%)
300	0,5	0,091 (+62,81%)
500	0,5	0,092 (+65.00%)
1000	0,5	0,093 (+67,66%)
1500	0,5	0,094 (+68,29%)
2000	0,5	0,093 (+66.52%)
3000	0,5	0,091 (+64,09%)
300	1,0	0,092 (+66,07%)
500	1,0	0,093 (+68,09%)
1000	1,0	0,095 (+71,31%)
1500	1,0	0,096 (+73,33%)
2000	1,0	0,097 (+73,67%)
3000	1,0	0,095 (+70,94%)

Pour réaliser les expérimentations, nous avons utilisé le système X-IOTA, avec un pré-calcul des relations entre doxels, et un calcul de propagation réalisé en PERL sur les fichiers de vecteur de X-IOTA en XML. De cette manière nous n'avons pas eu à modifier le SRI X-IOTA. Nous avons utilisé ce système car il permet d'accéder de manière simple aux vecteurs d'indexation, car ils sont contenus dans des fichiers au format XML, facilement modifiables par des programmes externes (comme ceux écrits pour l'occasion en PERL). Sur la collection d'INEX 2009, l'ordre de grandeur pour le calcul de toutes les relations, ainsi que tous les contextes est de 14 heures[4] , sur un cluster de 10 serveurs Unix (soit 140h sur un seul serveur). Le temps d'indexation et l'interrogation est du même ordre de grandeur avec et sans propagation. La taille des fichiers de descriptions (format texte) de tous les liens est entre 200Mo (ss2) et

4. Nous pensons à postériori que c'est une erreur d'avoir programmé ces modules en PERL, le temps important est probablement dû à une mauvaise gestion de la mémoire centrale par ce langage. Nous envisageons une nouvelle version en C++.

2Go (sec) pour un nombre de relations calculées entre 4 millions (ss2) et 30 millions (sec). La taille des contextes (fichiers au format XML) varie de 5Go (ss2), à 90Go (sec), ce qui peut poser un problème technique s'ils sont chargés en mémoire centrale, lors de leur utilisation à l'indexation. Nous avons résolu spécifiquement ce problème technique en réduisant les fichiers des contextes aux termes effectivement utilisés dans les requêtes. Cette simplification qui divise par 10 la taille des fichiers, est raisonnable dans le cas d'une évaluation de notre méthode avec des requêtes connues, et ajoute quelques heures supplémentaires pour ce filtrage.

7. Conclusion

Dans cet article, nous avons défini le contexte documentaire d'un élément de document structuré. Nous avons ensuite proposé de modéliser la recherche de tels doxels en contexte, par des modèles de langue. Notre proposition montre que la prise en compte de contextes documentaires dans des modèles de langues utilisant un lissage de Dirichlet peut être implanté sur un système de recherche d'information de manière efficace et sans une remise en cause important du processus d'indexation et de recherche.

Nous avons ensuite proposé de prendre en compte plusieurs contextes documentaires : l'un sur des distances dans les arbres de composition des document XML, l'autre en se basant sur des similarités. Des expérimentations nombreuses, portant sur des contextes limités à des doxels de même type, faisant varier les différents paramètres du modèle ont été menées et présentées sur le corpus INEX 2009. Les résultats obtenus sur 3 types de doxels, ss1, ss2 et sec, ont montré dans ces trois cas des améliorations relatives jusqu'à 73% par rapport à la non-utilisation des contextes, ce qui prouve de manière pratique le bénéfice de la prise en compte du contexte documentaire sur la collection INEX 2009. Il faut tout de même noter le coût induit par le calcul systématique de toutes les relations potentielles entre doxels, qui est de l'ordre de $O(n^2/2)$, avec n le nombre de doxel. Ce coût est prohibitif, lorsque l'on considère la collection complète de tous les doxels, il reste raisonnable et calculable lorsque l'on se limite aux doxels d'un même document. Dans ce cas, le coût est $O(m^2/2)$ avec m le nombre moyen de doxels par document qui est très inférieur à n.

Il existe de nombreuses directions à explorer. Au niveau expérimental, nous pouvons explorer davantage les espaces de paramètres, d'autres types de doxels, d'autres relations, utiliser d'autres collections que celle d'INEX qui est assez particulière car elle est basée sur l'encyclopédie libre Wikipédia. Au niveau théorique, nous pensons continuer à étudier les propriétés des formules utilisées, pour tenter d'intégrer des types d'utilisateurs et/ou de requêtes, et travailler sur un meilleur contrôle des différents paramètres.

8. Bibliographie

Arvola P., Kekäläinen J., Junkkari M., « Contextualization models for XML retrieval », *Information Processing & Management*, vol. 47, n° 5, p. 762 - 776, 2011.

Beigbeder M., « 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 of *Lecture Notes in Computer Science*, Springer Berlin / Heidelberg, p. 200-212, 2007.

Geva S., Kamps J., Trotman A., « 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 », *INEX*, vol. 6203 of *Lecture Notes in Computer Science*, Springer, 2010.

Ingwersen P., Järvelin K., « Information retrieval in context : IRiX », *SIGIR Forum*, vol. 39, p. 31-39, December, 2005.

Liu X., Croft W. B., « Cluster-based retrieval using language models », *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '04, ACM, New York, NY, USA, p. 186-193, 2004.

Lv Y., Zhai C., « Positional language models for information retrieval », *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '09, ACM, New York, NY, USA, p. 299-306, 2009.

Mulhem P., Chevallet J.-P., « Use of Language Model, Phrases and Wikipedia Forward Links for INEX 2009 », *INEX*, p. 103-111, 2009.

Myaeng S. H., Jang D.-H., Kim M.-S., Zhoo Z.-C., « A flexible model for retrieval of SGML documents », *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '98, ACM, New York, NY, USA, p. 138-145, 1998.

Ogilvie P., Callan J., « Hierarchical Language Models for XML Component Retrieval », *in* , N. Fuhr, , M. Lalmas, , S. Malik, , Z. Szlávik (eds), *Advances in XML Information Retrieval*, vol. 3493 of *Lecture Notes in Computer Science*, Springer Berlin / Heidelberg, p. 269-285, 2005.

Pinel-Sauvagnat K., Boughanem M., Chrisment C., « 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*, LNCS, Springer, http ://www.springerlink.com, p. 242-254, octobre, 2004. TauxAcceptation : (t :=123, l=26, c=18), (t= 35.7

Ponte J. M., Croft W. B., « A language modeling approach to information retrieval », *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '98, ACM, New York, NY, USA, p. 275-281, 1998.

Radhouani S., Chevallet J.-P., Géry M., « Un modèle à base de chemin de lecture pour la Recherche d'Informations précises sur le Web », *CORIA*, p. 249-270, 2004.

Shakery A., Zhai C., « Smoothing document language models with probabilistic term count propagation », *Inf. Retr.*, vol. 11, p. 139-164, April, 2008.

Smucker M. D., Allan J., An Investigation of Dirichlet Prior Smoothing's Performance Advantage, Technical Report n° CIIR Technical Report IR-548, University of Massachusetts, November, 2006.

Song F., Croft W. B., « A general language model for information retrieval », *Proceedings of the eighth international conference on Information and knowledge management*, CIKM '99, ACM, New York, NY, USA, p. 316-321, 1999.

Torjmen M., Pinel-Sauvagnat K., Boughanem M., « 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, octobre, 2010.

Zhai C., Lafferty J., « A study of smoothing methods for language models applied to Ad Hoc information retrieval », *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '01, ACM, New York, NY, USA, p. 334-342, 2001.