Modèles de langue appliqués à la recherche d'information contextuelle

Hugues Bouchard — Jian-Yun Nie

Laboratoire de recherche appliquée en linguistique information (RALI)
Département de recherche opérationnelle (DIRO), Université de Montréal
C.P. 6128 succursale Centre-ville, Montréal, Québec
Canada, H3C 3J7
{bouchahu, nie}@iro.umontreal.ca

RÉSUMÉ. *Il est reconnu que le contexte joue un rôle important en recherche d'information (RI). Or, très peu de systèmes opérationnels le considèrent. Dans cet article, nous considérons un des aspects du contexte – le domaine d'intérêt de l'usager. Nous caractérisons un domaine d'intérêt par un ensemble de documents. Nous utilisons une approche de modélisation de langue statistique pour établir un modèle de langue du domaine. Ce modèle est utilisé de trois façons : pour étendre la requête initiale, pour réordonner les documents retrouvés, et pour exploiter les relations lexicales spécifiques au domaine. Nos expériences montrent que nos modèles, qui tiennent compte du contexte, surpassent en performance le modèle traditionnel. Ces expérimentations montrent que le contexte doit être pris en compte si on veut améliorer la performance des systèmes de RI actuels.*

ABSTRACT. *It is recognized that context plays a crucial role in information retrieval (IR). However, few operational systems take it into account. In this paper, we consider one of the important aspects of the context – the domain of interest of the user. We assume that each domain can be characterized by a set of documents. This latter is used to build a statistical language model, which is then integrated in three different ways: to expand the original query model, to re-rank the retrieved documents and to extract lexical dependencies of the domain. Our experiments show that our contextual approaches improve significantly the retrieval performance compared to the basic approach, which does not consider context. These experiments provide evidence that context should be taken into account to further improve the effectiveness of current IR systems.*

MOTS-CLÉS : *modèle de langue, recherche d'information, contexte.*

KEYWORDS: *language model, information retrieval, context.*

1. Introduction

Les approches existantes en recherche d'information (RI) se basent sur une similarité entre la requête et les documents de la collection, sans prendre en compte le contexte de la recherche (Belkin, 1993). Malgré le progrès fait pour représenter et comparer la requête et les documents, ces approches semblent avoir atteint leurs limites, et une amélioration supplémentaire requiert la prise en considération du contexte de la recherche (Allan et *al*, 2002).

La requête adressée au système est généralement très brève, 2 à 4 mots en moyenne. Elle est donc une expression très partielle et souvent ambigüe du besoin d'information de l'usager. Considérée isolément, elle ne suffit pas à identifier clairement ce que l'usager recherche. En tenant compte du contexte, l'information partielle de la requête peut être complétée et l'ambiguïté peut être résolue à un certain degré. Par exemple, la requête « java bean » est ambiguë : elle peut désigner un langage de programmation ou des grains de café. Or, sachant que l'utilisateur s'est beaucoup intéressé aux documents en informatique dans le passé, l'ambiguïté peut être levée. Dans la pratique, sans procéder à une véritable désambiguïsation, le contexte peut aider à préciser la requête en y ajoutant les termes importants du domaine, comme par exemple « programme », « package », etc. Ce faisant, les documents concernant le langage de programmation seront favorisés.

L'importance du contexte est reconnu en RI depuis longtemps (Ingwersen et al, 2004). Pourtant, il existe peu de modèles opérationnels qui l'intègrent dans l'opération de recherche. La notion même de contexte demeure floue à ce jour, et est utilisée pour désigner divers aspects, allant du contexte physique (matériel) au contexte cognitif. Plusieurs études en sciences de l'information ont toutefois montré que l'usager, et surtout son domaine d'expertise, est un des facteurs contextuels les plus importants (Park, 1994 ; Morris, 1994). Dans cet article, nous tenons compte de cet aspect important du contexte: le domaine d'intérêt de l'usager. Cette étude est effectuée dans le cadre suivant : on suppose que l'usager peut indiquer le domaine de son besoin d'information, en plus de formuler une requête. Un domaine est choisi parmi un ensemble de domaines préétablis (par exemple, « Commerce International », « Informatique », etc.). Dans cet article, nous utilisons une approche de modèle de langue statistique pour modéliser le domaine. Le modèle de langue inféré des documents du domaine sera utilisé de différentes façons pour compléter la requête. Nos expériences montrent que la prise en compte du contexte (le domaine d'intérêt de l'usager) nous permet d'obtenir de grandes améliorations par rapport à une approche non contextuelle. Ceci montre de façon concrète que la prise en compte du contexte est une voie à suivre pour améliorer les systèmes de RI actuels.

2. La recherche d'information contextuelle

Depuis quelques années, on voit apparaître en RI des approches qui considèrent certains aspects du contexte. Par exemple, dans (Lau et *al*, 2004), le contexte cognitif de l'usager est modélisé par un ensemble de croyances. Par une approche logique, l'interprétation de la requête la plus compatible avec les croyances de l'usager est retenue. Remarquons qu'en pratique, il est difficile d'extraire les relations logiques strictes. Ainsi, d'autres approches utilisent une modélisation plus flexible du domaine.

Il est possible de contextualiser une requête en exploitant des évidences implicites inférées du comportement de l'usager (par rétroaction de pertinence). Plusieurs facteurs témoignent de l'appréciation de l'usager (Kelly et al, 2003), comme la consultation, la sauvegarde et le mouvement oculaire. Dans (Shen et *al*, 2005), la requête est redéfinie d'après les documents consultés durant la session. Cependant, le profil est seulement utilisé pour une session.

Une autre approche est d'utiliser le domaine d'intérêt de l'usager pour réordonner les documents retrouvés selon leur correspondance au domaine. Un domaine d'intérêt peut être représenté de différentes manières. Dans (Chirita et *al*, 2005), une hiérarchie des domaines préconisés par l'usager est utilisée. (Kim et *al*, 2005) considèrent plutôt l'ensemble des documents consultés antérieurement par l'usager comme une spécification d'un domaine. Dans (Teevan et *al*, 2005), on infère un modèle probabiliste de l'ensemble des documents présents dans l'ordinateur de l'usager. Dans tous ces travaux, l'idée est de réordonner les documents retrouvés selon leur degré d'appartenance au domaine préconisé par l'usager.

Dans cette étude, nous considérons qu'un domaine d'intérêt est spécifié par un ensemble de documents. Un modèle de langue est construit pour chaque domaine. En spécifiant un domaine d'intérêt pour sa recherche, l'usager peut aider à préciser sa requête. Comme nous utilisons une approche de modélisation de langue statistique, nous allons en faire une description dans la section suivante.

3. Modèles de langue

Le principe de base des modèles de langue en RI est d'ordonner chaque document D de la collection C suivant leur capacité à générer la requête Q. Ainsi, il s'agit d'estimer la probabilité de génération P(Q|D) comme suit :

$$P(Q \mid D) = \prod_{t \in Q} P(t \mid \theta_D)^{c(t;Q)} \qquad [1]$$

où c(t ;Q) est la fréquence du terme t dans la requête Q, et θ_D est le modèle du document.

Suivant le cadre de travail proposé dans (Lafferty et al, 2001), la similarité entre un document et une requête peut aussi être exprimée par la mesure de divergence de Kullback-Leibler. La KL-divergence exprime la distance entre les distributions mises en relation. Ainsi, une fonction d'ordonnancement peut être définie comme suit :

$$Score(Q,D) = -KL(\theta_Q \parallel \theta_D) = \sum_t P(t \mid \theta_Q) \log \frac{P(t \mid \theta_D)}{P(t \mid \theta_Q)} \propto \sum_t P(t \mid \theta_Q) \log P(t \mid \theta_D) \qquad [2]$$

où θ_Q est le modèle de Q, généralement estimé par fréquence relative des mots-clefs dans la requête. Cette fonction de score peut aussi être vue comme une entropie croisée.

Afin d'éviter une probabilité nulle, ainsi que pour modéliser la spécificité des termes de la requête, il s'avère essentiel de lisser le modèle du document par un modèle de retrait, généralement le modèle de la collection (Zhai et *al*, 2001). Une des stratégies fréquemment utilisées est de lisser le modèle du document par le lissage de Jelinek-Mercer, soit : $P(t \mid \theta'_D) = (1-\lambda) P(t \mid \theta_D) + \lambda P(t \mid \theta_C)$.

4. Contexte usager

Nous envisageons le scénario d'utilisation suivant pour l'identification d'un domaine particulier : pour chaque requête émise, l'usager identifie un domaine pour son besoin d'information, soit l'une des catégories de la hiérarchie préétablie, par exemple « Histoire », « Informatique », etc. Ce scénario est réaliste. En effet, l'identification d'un domaine par l'usager n'augmente pas substantiellement la charge cognitive de l'usager dans la mesure où il y a un choix restreint de domaines.

À chaque domaine correspond un ensemble de documents du domaine. Ainsi, ces documents caractérisent le domaine d'une façon implicite. Dans nos travaux, ces documents sont récupérés suivant deux approches. La première approche consiste à récupérer tous les documents que les usagers du système ont jugés pertinents pour des requêtes émises antérieurement dans le domaine. L'alternative est d'utiliser une hiérarchie de classes existantes, par exemple ODP [1] pour retrouver un ensemble de documents spécifiques à chaque domaine. Ces deux approches seront explorées dans nos expérimentations.

Les documents du domaine contiennent des éléments caractéristiques du domaine, c'est-à-dire les termes importants du domaine, mais aussi beaucoup d'autres éléments généraux de la langue qui ne sont pas informatifs (du bruit). Il est important de dégager les éléments significatifs par un moyen résistant au bruit. Les modèles de langue sont appropriés pour jouer ce rôle. Nous proposons donc d'établir un modèle de langue pour chaque domaine.

[1]. The Open Directory Project, http://dmoz.org/about.html

Nous considérons que chacun des documents du domaine est généré conjointement par un modèle du domaine spécifique et un modèle plus général :

$$P(D \mid \theta'_{Dom}) = \prod_{t \in D} [(1-\eta)P(t \mid \theta_{Dom}) + \eta P(t \mid \theta_C)]^{c(t;D)} \quad [3]$$

où θ_C est le modèle de la collection utilisé pour approximer le modèle général, θ_{Dom} est le modèle du domaine spécifique que l'on cherche à extraire, et θ'_{Dom} est le modèle engendré par les deux sources. Afin d'estimer θ_{Dom}, nous utilisons un algorithme EM (Zhai et *al*, 2001), qui détermine θ_{Dom} de façon à maximiser $P(Dom| \theta'_{Dom})$, où Dom est l'ensemble des documents du domaine, c'est-à-dire :

$$\arg\max_{\theta_{Dom}} P(Dom \mid \theta'_{Dom}) = \prod_{D \in Dom} \prod_{t \in D} [(1-\eta)P(t \mid \theta_{Dom}) + \eta P(t \mid \theta_C)]^{c(t;D)} \quad [4]$$

L'itération de l'algorithme EM calcule les deux paramètres suivants :

$$w^{(i)}(t) = \frac{(1-\eta)P(t \mid \theta_{Dom})}{(1-\eta)P(t \mid \theta_{Dom}) + \eta P(t \mid \theta_C)} \quad \text{(E-Étape)} \quad [5]$$

$$P^{(i+1)}(t \mid \theta_{Dom}) = \frac{\sum_{D \in Dom} c(t;D) w^{(i)}(t)}{\sum_{t'} \sum_{D \in Dom} c(t';D) w^{(i)}(t')} \quad \text{(M-Étape)} \quad [6]$$

Essentiellement, en estimant θ_{Dom} de cette façon, nous tentons de purifier le modèle du domaine du bruit présent dans les documents. Au terme des itérations, il en résulte un modèle de langue dont la probabilité de masse se concentre sur les termes significatifs du domaine, peu observés dans le modèle général.

5. Modèles de recherche d'information contextuelle

Nous présentons ici trois méthodes d'intégration du modèle du domaine.

5.1. Compléter le modèle de requête

Ayant un modèle qui caractérise le domaine de la requête, le modèle de langue de la requête peut être combiné au modèle du domaine avec le lissage suivant :

$$\theta'_Q = (1-\alpha)\theta_Q + \alpha\theta_{Dom} \quad [7]$$

où θ'_Q est le nouveau modèle engendré. Comme nous l'avons mentionné plus tôt, la rétroaction de pertinence peut aussi être utilisée pour compléter la requête. Un modèle θ_R est créé à partir des n premiers documents retrouvés, et est utilisé pour compléter la requête comme suit :

$$\theta'_Q = (1 - \alpha - \beta)\theta_Q + \alpha\theta_{Dom} + \beta\theta_R \quad [8]$$

Tout comme θ_{Dom}, θ_R est estimé en appliquant l'algorithme EM décrit à la section 4 sur les documents retenus. Une fois le nouveau modèle de la requête θ'_Q obtenu, le score d'un document D pour une requête Q du domaine Dom est déterminé comme suit :

$$Score(Q, D, Dom) = -KL(\theta'_Q \| \theta'_D) \propto \sum_t P(t \mid \theta'_Q) \log P(t \mid \theta'_D) \quad [9]$$

où θ'_D est le modèle lissé du document tel que décrit à la section 3.

5.2. Ré-ordonnancement des documents

Une seconde stratégie est d'utiliser le modèle du domaine pour réordonner les documents retrouvés, comme dans (Chirita et *al*, 2005 ; Kim et *al*, 2005 ; Teevan et *al*, 2005). Les documents retrouvés avec la requête sont réordonnés d'après leur correspondance avec le domaine. Cette approche permet de favoriser les documents qui s'apparentent les plus au domaine de la requête. Nous utilisons la fonction suivante pour combiner la similarité initiale avec la correspondance au domaine :

$$Score(Q, D, Dom) = -\left[(1 - \chi)KL(\theta_Q \| \theta'_D) + \chi KL(\theta_D \| \theta'_{Dom})\right] \quad [10]$$

Le coefficient χ contrôle l'importance accordée au domaine lors de l'estimation de la pertinence d'un document.

5.3. Exploiter les dépendances lexicales du domaine

L'ensemble des documents du domaine ne fournissent pas seulement un vocabulaire spécifique, mais aussi des relations particulières entre les termes. Par exemple, dans le domaine « Finance », le terme « budget » est fortement relié au terme « planification ». Ainsi, dans ce troisième modèle, on exploite les dépendances lexicales du domaine. Ces dépendances lexicales sont inférées de l'ensemble des documents du domaine en utilisant les cooccurrences existant entre les termes observés. Inspiré du cadre de travail proposé par (Berger et *al*, 1999), l'idée est d'étendre la requête avec tous les termes fortement corrélés aux mots-clefs qui la composent. Intuitivement, il s'agit de lisser le modèle du document non pas uniformément, mais plutôt sémantiquement suivant les dépendances lexicales entre les termes du domaine. Étant donné une requête Q et le domaine préconisé Dom, la probabilité d'observer le document D peut être approximée comme suit :

$$P(D \mid Q, Dom) = \frac{P(Q \mid D, Dom)P(D \mid Dom)}{P(Q \mid Dom)} \propto P(Q \mid D, Dom)P(D \mid Dom) \quad [11]$$

La probabilité P(D|Dom) peut être calculée comme une probabilité de génération :

$$P(D \mid Dom) = \prod_{t \in D} P(t \mid \theta'_{Dom})^{c(t;D)} \quad [12]$$

La probabilité P(Q|D,Dom) traduit la probabilité de générer la requête étant donné le document et les dépendances lexicales du domaine. Le modèle combinant D et Dom est dénoté par le modèle de dépendance $\Phi_{D,Dom}$. Suivant (Berger et *al*, 1999), on a :

$$P(t \mid \Phi_{D,Dom}) = \sum_{v \in D} t_{Dom}(t \mid v) P(v \mid \theta'_D) \quad [13]$$

où $P(v \mid \theta'_D)$ est un modèle lissé du document, et $t_{Dom}(t|v)$ est la probabilité de dépendance de t à v estimée dans Dom. Cette probabilité de dépendance est basée sur la cooccurrence des termes dans les documents du domaine. Nous considérons une fenêtre de longueur 5 pour la cooccurrence. Soit c(<t,v> ;D), le nombre de fois que t occurre avec v dans un document D de Dom. Nous définissons la dépendance lexicale comme suit :

$$t_{Dom}(t \mid v) = \frac{\sum_{D \in Dom} c(\langle t, v \rangle; D)}{\sum_{t'} \sum_{D \in Dom} c(\langle t'.v \rangle; D)} \quad [14]$$

Étant donné le faible recouvrement des dépendances lexicales du domaine, nous devons considérer également les dépendances lexicales de la collection, qui sont intégrées par le lissage suivant :

$$P(t \mid \Phi'_{D,Dom}) = \sum_{v \in D} [(1-\mu) t_{Dom}(t \mid v) + \mu t_c(t \mid v)] P(t \mid \theta'_D) \quad [15]$$

Finalement, le modèle de dépendance $\Phi'_{D,Dom}$ doit aussi être interpolé au modèle uni-gramme du document θ'_D. Ainsi :

$$P(Q \mid D, Dom) = \prod_{t \in Q} P(t \mid \Phi''_{D,Dom})^{c(t;Q)} = \prod_{t \in Q} [(1-\gamma) P(t \mid \Phi'_{D,Dom}) + \gamma P(t \mid \theta'_D)]^{c(t;Q)} \quad [16]$$

Substituant les équations [12] et [16] dans l'expression logarithmique de [11], on obtient :

$$\log P(D \mid Q, Dom) = \sum_{t \in Q} c(r;Q) \log P(t \mid \Phi''_{D,Dom}) + \sum_{t \in D} c(t;D) \log P(t \mid \theta'_{Dom})$$

$$\propto |Q| \sum_{t \in Q} P(t \mid \theta_Q) \log P(t \mid \Phi''_{D,Dom}) + |D| \sum_{t \in D} P(t \mid \theta_D) \log P(t \mid \theta'_{Dom}) \quad [17]$$

Les deux composantes de l'équation [17] sont en fait $KL(\theta_Q \| \Phi''_{D,Dom})$ et $KL(\theta_D \| \theta'_{Dom})$ multipliés par deux constantes reliées à la requête et au document. Pour simplifier, on suppose qu'elles sont invariables à travers la collection. On les dénote par (1-δ) et δ. Ainsi :

$$Score(Q, D, Dom) = -\left[(1-\delta) KL(\theta_Q \| \Phi''_{D,Dom}) + \delta KL(\theta_D \| \theta'_{Dom})\right] \quad [18]$$

6. Expériences et résultats

Nos expériences ont été réalisées en utilisant les outils de Lemur [2]. Deux collections ad hoc de TREC [3] ont été considérées, soient les documents des disques 1-2, et ceux du disque 3. Nous avons utilisé seulement les titres des requêtes 51 à 150. Ces requêtes ont la particularité de posséder un champ indiquant le domaine dans lequel elles sont exprimées. Ainsi, il est possible de simuler nos approches en exploitant l'ensemble des documents caractérisant le domaine associé à la requête. Les requêtes sont distribuées sur 13 domaines, tels que « Environnement », « Finance », « Militaire », « Science et Technologie », etc. Nous utilisons l'algorithme de Krovetz (Krovetz, 1993) pour la lemmatisation. Les termes fonctionnels sont retirés.

Suivant les deux approches mentionnées à la section 3, les documents caractérisant le domaine sont récupérés des collections de TREC et de l'annuaire web d'ODP. Pour les documents issus des collections de TREC, nous considérons les documents jugés pertinents pour les requêtes 51 à 150 du domaine. Afin d'éviter un biais favorisant nos modèles, aucun document de la collection test ni aucun document jugé pertinent pour la requête en cours n'est présent dans le domaine. Pour les documents issus des répertoires d'ODP, nous avons récupéré les documents dont la catégorie s'apparente à celle du domaine choisi. Les correspondances entre les domaines et les catégories d'ODP ont été établies manuellement. Par exemple, pour le domaine « Environnement », les catégories suivantes ont été retenues : « Science/Environnement », « Science/ Biologie/Écologie » et « Société/Problèmes et Débats/Nature et Environnement ». Ici, nous avons créé ces correspondances manuellement. Il serait possible de le faire d'une façon automatique à l'avenir, en utilisant les techniques de classification.

L'objectif de nos expérimentations n'est pas de montrer que nos approches sont les plus performantes en RI contextuelle, mais bien qu'il est avantageux de considérer le domaine de l'usager d'après les façons proposées. Chacun des

[2]. The Lemur Toolkit: http://www.lemurproject.org
[3]. Text Retrieval Conference: http://trec.nist.gov

paramètres caractérisant nos modèles ont été fixés au terme d'une exploration discrète de l'espace des paramètres. Dans la mesure où leurs valeurs optimales ne semblent pas dépendre du jeu de données, nous avons pu fixer les paramètres sans risque de sur-apprentissage. Par ailleurs, il s'avère que seuls un nombre restreint de paramètres ont une influence directe sur les performances obtenues, soient le coefficient de bruit η dans l'ensemble du domaine, les coefficients de lissages du modèle de la requête α et β, et les coefficients d'interpolations de similarité χ et δ. Ainsi, dans la pratique, seuls quelques paramètres doivent êtres évalués avec soin.

Afin d'évaluer la contribution du modèle du domaine au modèle de la requête, nous avons comparé les performances du modèle contextuel de l'équation [7] et [9] ($\theta_Q+\theta_{Dom}$) à celles d'un modèle traditionnel définit par l'équation [2] (θ_Q). Les performances du modèle de l'équation [8] et [9] ($\theta_Q+\theta_{Dom}+\theta_R$) ont aussi été comparées à celles d'un modèle traditionnel d'expansion de requête ($\theta_Q+\theta_R$) utilisant la rétroaction de pertinence. Les 20 premiers documents de la collection jugés les plus pertinents pour la requête selon l'équation [1] ont été considérés pour la rétroaction de pertinence. Par ailleurs, afin d'obtenir un modèle du domaine plus centré sur la requête, dans la méthode $\theta_Q+\theta_{Dom}+\theta_R$, seuls les 20 premiers documents de l'ensemble du domaine jugés les plus pertinents pour la requête selon l'équation [1] ont été considérés.

Dans la figure 1, nous pouvons observer une amélioration de 13% à 19% de la précision moyenne lorsque le modèle du domaine est interpolé au modèle original de la requête ($\theta_Q+\theta_{Dom}$). Cette comparaison montre clairement que le modèle du domaine permet de compléter l'information partielle transmise par la requête.

Collection	Mesure	Θ_Q	$\Theta_Q + \Theta_{Dom}$		$\Theta_Q + \Theta_R$	$\Theta_Q + \Theta_{Dom} + \Theta_R$	
TREC Disques 1 & 2	Précision	0.1867	0.2115 (+13%)	0.2206 (+18%)	0.2701	0.2899 (+7%)	0.2902 (+7%)
	Rappel (28 030)	12 290	13 464 (+4%)	14 039 (+6%)	15 117	15 990 (+6%)	16 079 (+6%)
TREC Disque 3	Précision	0.1770	0.2112 (+19%)	0.2087 (+18%)	0.2462	0.2605 (+6%)	0.2599 (+6%)
	Rappel (20 334)	8 634	9 784 (+6%)	10 058 (+7%)	10 661	11 056 (+4%)	11 080 (+4%)
			Profil TREC	Profil ODP		Profil TREC	Profil ODP

Figure 1. *Combinaison du modèle original de la requête θ_Q à un modèle du domaine θ_{Dom} et/ou un modèle de rétroaction de pertinence θ_R.*

Les performances de la colonne $\theta_Q+\theta_R$ montre que le modèle de rétroaction est plus apte à expliciter le besoin de l'usager que le modèle du domaine. Malgré ce fait, il demeure intéressant de constater (colonne $\theta_Q+\theta_{Dom}+\theta_R$) que le modèle du domaine peut apporter une amélioration supplémentaire à ce dernier. Bien que la précision moyenne ne soit améliorée que 6%-7%, le modèle du domaine exprime manifestement des informations que le modèle de rétroaction n'exprime pas.

Il est intéressant d'observer que le modèle du domaine apporte la même contribution qu'il soit défini avec les documents récupérés des collections TREC ou avec les documents d'ODP. En regard aux performances mesurées, les deux sources semblent équivalentes. Ceci montre que les deux façons de générer le modèle du domaine sont aussi valables l'une que l'autre. En pratique, il est possible d'utiliser une hiérarchie de classes disponibles pour estimer des modèles de domaines et de les utiliser pour la recherche sur une collection indépendante.

Collection	Mesure	Modèle de base	Modèle de ré-ordonnancement	
TREC Disques 1 & 2	Précision	0.1867	0.2135 (+14%)	0.2108 (+13%)
	Rappel (28 030)	12 290	13 347 (+4%)	13 230 (+3%)
TREC Disque 3	Précision	0.1770	0.1922 (+8%)	0.1893 (+7%)
	Rappel (20 334)	8 634	9 170 (+3%)	9 144 (+3%)
			Profil TREC	Profil ODP

Figure 2. *Ré-ordonnancement des documents initialement retrouvés selon leur degré d'appartenance au domaine.*

Dans la seconde expérimentation, nous comparons les performances du modèle de ré-ordonnancement de l'équation [10] à celles d'un modèle de base défini en [2]. Dans la figure 2, nous constatons qu'il est avantageux de procéder de cette façon. Bien que les gains soient moindres que ceux obtenus avec la première approche (voir figure 1), il y a de fortes améliorations de la précision moyenne (7% à 14%).

La dernière série d'expérimentations porte sur les dépendances lexicales du domaine. Le modèle défini à l'équation [18] est comparé à un modèle similaire qui ne tient compte que des dépendances lexicales inférées de la collection. En utilisant les dépendances lexicales extraites de toute la collection, nous pouvons observer dans la figure 3 qu'aucune amélioration n'est observée par rapport au modèle de base (voir figure 2). Toutefois, en utilisant les dépendances lexicales du domaine, nous obtenons des améliorations significatives de 11% à 18%.

Collection	Mesure	Modèle de dépendances lexicales de la collection	Modèle de dépendances lexicales du domaine	
TREC Disques 1 & 2	Précision	0.1837	0.2162 (+18%)	0.2170 (+18%)
	Rappel (28 030)	12 321	13 378 (+4%)	13 395 (+4%)
TREC Disque 3	Précision	0.1771	0.1957 (+11%)	0.2170 (+18%)
	Rappel (20 334)	8 975	9 439 (+2%)	13 395 (+4%)
			Profil TREC	Profil ODP

Figure 3. *Performances obtenues en exploitant les dépendances lexicales du domaine vs les dépendances lexicales de la collection.*

À travers nos expérimentations, nous pouvons voir que les trois façons d'intégrer le modèle du domaine aboutissent à une augmentation sensible des performances. Ces séries d'expérimentations montrent l'importance de tenir compte du domaine. Elles témoignent aussi de la validité des méthodes que nous avons proposées dans cet article.

Les approches que nous avons proposées ici sont réalisables car la plupart des calculs requis pour représenter et intégrer le domaine d'intérêt de l'usager peuvent être faits hors ligne. Ainsi, le temps en ligne pour l'évaluation d'une requête demeure rapide. Le modèle de dépendance nécessite toutefois beaucoup de mémoire pour emmagasiner les probabilités de dépendances entre les termes (1 Go).

7. Conclusion

Dans ce papier, nous avons présentés trois modèles opérationnels qui exploitent un aspect du contexte cognitif de l'usager, son domaine d'intérêt. Caractérisé initialement par un ensemble de documents, le domaine est représenté par un modèle de langue. Le modèle du domaine est considéré pour compléter le modèle initial de la requête, pour réordonner les documents préalablement retrouvés et pour exploiter les dépendances lexicales du domaine.

Nos expériences ont montré qu'il est avantageux de considérer le domaine d'intérêt de l'usager. Quand un modèle du domaine est utilisé, nous avons observé de forte amélioration en performance. Il est intéressant de constater que l'approche la plus performante est celle qui utilise le domaine pour compléter la requête (première approche). Cela suggère que la plus grande faiblesse de la requête, qui est liée à sa taille réduite, peut être compensée par la prise en compte du domaine. Ainsi, le domaine d'intérêt de l'usager est un élément important pour mieux comprendre et évaluer la requête.

Les modèles que nous avons proposés dans cet article ne doivent pas être utilisés pour remplacer d'autres méthodes efficaces, mais plutôt pour apporter des éléments complémentaires. En effet, le modèle du domaine peut être combiné à la méthode de pseudo-rétroaction de pertinence, et ceci a produit une amélioration supplémentaire.

Cette étude est encore préliminaire, et plusieurs aspects peuvent être améliorés. Nous avons considéré tous les domaines de la même façon. Mais dans les faits, les domaines associés aux requêtes ne sont pas égaux en ce qui concerne la spécificité ou l'homogénéité. Certains domaines sont plus spécifiques que d'autres. Par exemple, le domaine « Militaire » s'avère plus homogène que le domaine « Science et Technologie ». Ainsi, il serait intéressant de considérer la spécificité ou l'homogénéité du domaine dans son utilisation. Cela sera un aspect que nous allons étudier dans le futur. Par ailleurs, il serait intéressant de vérifier si l'ensemble des documents du domaine peuvent être déterminés au moyen d'une classification automatique tout en offrant les mêmes performances.

8. Bibliographies

Allan J., Aslam J., Belkin N.J., Buckley C., « Challenges in information retrieval and language modeling », *Communications of the ACM*, vol. 37, no 1, 2003, p.31-47.

Belkin, N.J., « Interaction with texts: Information retrieval as information-seeking behaviour », *Proceedings of the Information Retrieval '93 conference*, 1993, p.55-66.

Berger A., Lafferty J., « Information retrieval as statistical translation », *Proceedings of the ACM SIGIR '99 conference*, 1999, p. 222-229.

Chirita P., Nejdl W., Paiu R., Kohlsch C., « Using ODP metadata to personalize search », *Proceedings of the ACM SIGIR '05 conference*, 2005, p. 178-185.

Ingwersen, P., Belkin, N., « Information retrieval in context - IRIX workshop at SIGIR 2004 », SIGIR Forum, vol. 38, no 2, 2004, p.50-52.

Kelly, D, Teevan, J., « Implicit feedback for inferring user preference », SIGIR forum, vol. 37, no 2, 2003, p.18-28.

Kim H., Chan P., Personalized ranking of search results with implicitly learned user interests hierarchies, Computer Sciences Department Technical Report no CS-2005-11, Florida Institute of Technology, 2005.

Krovetz, R., « Viewing morphology as an inference process », *Proceeding of the ACM SIGIR '93 conference*, 1993, p.191-202.

Lafferty J., Zhai C., « Document language models, query models, and risk minimization for information retrieval », P*roceedings of the ACM SIGIR '01 conference*, 2001, p. 111-119.

Morris R., « Toward a user-centered information service », *Journal of the American society for information science*, vol. 45, no 1, 1994, p. 20-30.

Park T.K., « Toward a theory of user-based relevance: A call for a new paradigm of inquiry », *Journal of the American society for information science*, vol. 45, no 3, 1994, p. 135-141.

Shen X., Tan B., Zhai C., « Context-sensitive information retrieval using implicit feedback », *Proceedings of the ACM SIGIR '05 conference*, 2005, p. 43-50.

Teevan J., Dumais S., Horvitz E., « Personalizing search via automated analysis of interests and activities », P*roceedings of the ACM SIGIR '05 conference*, 2005, p. 449-456.

Zhai C., Lafferty J., « A study of smoothing methods for language models applied to ad hoc information retrieval », *Proceedings of the ACM SIGIR '01 conference*, 2001, p. 334-342.

Zhai C., Lafferty J., « Model-based feedback in the language modeling approach to information retrieval », *Proceedings of the CIKM '01 conference* , 2001, p. 403-410.