Modèles de langue appliqués à la recherche d'information contextuelle

Hugues Bouchard* et Jian-Yun Nie†

* RALI, Université de Montréal, bouchahu@iro.umontreal.ca
† RALI, Université de Montréal, nie@iro.umontreal.ca

Résumé

Il est reconnu que le contexte joue un rôle important en recherche d'information (RI). Or, peu de systèmes opérationnels en tiennent compte. Dans cet article, nous considérons un des aspects importants du contexte - le domaine d'intérêt de l'usager. La requête, généralement très courte, est interprétée dans son domaine particulier. Ceci se traduit par la création d'un nouveau modèle de langue pour la requête. Dans cette étude, le domaine est utilisée de trois façons différentes : pour compléter la requête, pour réordonner les résultats de recherche ou pour étendre la requête en utilisant les relations lexicales extraites du domaine. Nos expériences sur des collections TREC montrent clairement l'utilité du domaine. Nos approches d'intégration du domaine nous permettent d'améliorer les performances de recherche de façon significative.

Mots-clés *: modèle de langue, recherche d'information, contexte.*

Abstract

It is recognized that context plays a crucial role in information retrieval (IR). However, few operational IR systems take it into account. In this paper, we consider one of the important aspects of context - the domain of interest of the user. A query, which is usually short, is interpreted within its domain. This is implemented as the creation of a new statistical language model for the query. The domain is exploited in three different ways: to complete the query, to reorder the retrieval results, or to expand the query using lexical relations extracted from the domain. Our experiments on TREC collections demonstrate clearly the utility of domain for IR. The methods we suggested all lead to significant improvements of retrieval effectiveness.

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

1. INTRODUCTION

Traditionnellement, l'estimation de degré de pertinence d'un document à l'égard d'une requête se résume à une opération d'appariement entre leurs termes. La stratégie traditionnelle consiste à représenter la requête et le document par un vecteur selon les termes observés, et à mesurer la similarité entre eux. La littérature a souvent soulevé le fait que les termes utilisés dans une telle approche sont en effet supposés d'être indépendants, ce qui ne correspond pas à la réalité. Il a aussi été observé que cette façon d'évaluer le degré de pertinence ne considère que le document et la requête, et tous les autres éléments entourant la recherche sont ignorés [2]. Or, l'estimation de la pertinence d'un document pour une requête est fortement dépendante du contexte dans laquelle la recherche est effectuée, par exemple, le but de la recherche, les connaissances de l'usager, les informations déjà disponibles, etc. Tous ces facteurs contextuels jouent un rôle important dans le jugement sur la pertinence du document, et il est nécessaire d'en tenir compte afin d'améliorer la performance des systèmes de recherche d'information (RI) [1, 20].

Le contexte est toutefois une notion très vague qui dénote pratiquement tout élément qui peut affecter le jugement de la pertinence [27]. Parmi ces éléments, on retrouve la structure de similarité du corpus, l'environnement lexical, de même que l'usager lui-même (avec ses connaissances et ses préférences). L'importance du contexte s'explique par sa capacité à donner sens au phénomène observé. C'est effectivement en regard aux éléments environnants, aux conditions favorisant l'émergence de la requête et des documents que leur sens est révélé. Plusieurs études en sciences cognitives s'accordent dans ce sens : l'information doit être comprise dans son contexte [16, 22], et la prise en considération du contexte dans la RI est cruciale [17].

La prise en compte du contexte est d'autant plus nécessaire en RI que la requête adressée au système est généralement très brève, 2 à 4 mots en moyenne [24]. En effet, la requête n'est qu'une expression partielle et souvent ambiguë du besoin d'information de l'usager. Considérée isolément, elle ne suffit pas à identifier clairement ce que l'usager recherche. En considérant le 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 ambigüe. Sans regard au contexte dans lequel elle est exprimée, elle dénote autant un langage de programmation utilisé dans le développement d'applications Web, que les grains de café produits sur une île d'Indonésie. Toutefois, sachant que l'usager a exprimé son besoin dans le cadre de sa profession en informatique, l'ambiguïté peut être levée. Dans ce cas particulier, le contexte professionnel de l'usager aide à résoudre l'ambiguïté des termes de la requête. De façon générale, le contexte peut aider à préciser ou à compléter l'information partielle véhiculée par la requête. Une des façons pour résoudre l'ambiguïté consiste à déterminer le sens des mots dans la requête, ou à procéder à une désambiguïsation. Cependant, ceci est souvent très difficile [28]. Une autre façon moins rigide consiste à réduire l'ambiguïté de la requête en ajoutant d'autres mots reliés au contexte. Pour la requête « java bean », en ajoutant des termes caractéristique du domaine « Informatique », tels que « application », « api » et « programme », l'ambiguïté de la requête est de beaucoup réduite. Ceci constitue une des méthodes que nous allons utiliser dans notre étude.

Dans cet article, nous n'abordons pas tous les aspects du contexte, mais nous nous limitons au contexte cognitif de l'usager, plus précisément à son domaine d'intérêt. Nous entendons par « domaine d'intérêt » le champ d'expertise de l'usager, soit l'une des sphères du savoir et de la culture qui suscite son intérêt. Plusieurs études en sciences de l'information ont montré que l'usager est un des facteurs contextuels les plus importants [18, 19], et plus particulièrement le domaine d'intérêt de l'usager est celui qui affecte le plus le jugement de la pertinence parmi tous les facteurs contextuels [25]. Ainsi, le domaine d'intérêt de l'usager est le premier facteur à tenir compte dans une RI contextuelle.

Il y a différentes façons d'identifier le domaine d'intérêt de l'usager. Par exemple, l'usager peut créer un profil qui précise ses domaines d'intérêt. On peut aussi utiliser un système de classification automatique pour déterminer les domaines d'intérêts potentiels, en se basant sur les documents cherchés ou lus par l'usager et les requêtes posées antérieurement. L'identification du domaine d'intérêt n'est pas le sujet central de cette étude. Ainsi, dans cette étude, nous utilisons une approche simple : nous supposons qu'une requête n'est pas soumise toute seule, mais avec une indication de son domaine d'intérêt, tel que « Art Visuel », « Commerce International », « Environnement », « Informatique », etc. C'est d'ailleurs le cas dans TREC 1-2, où chaque sujet (*topic*) contient une telle identification du domaine d'intérêt. Notre objectif dans cet article est de montrer que le domaine d'intérêt, une fois identifié, peut beaucoup contribuer à améliorer la performance de recherche.

Pour exploiter le domaine d'intérêt de l'usager, deux stratégies peuvent être envisagées, soient : créer manuellement un modèle sémantique et conceptuel du domaine, ou bien créer automatiquement un modèle statistique du domaine. La première approche a été explorée dans certaines études antérieures. Dans [4, 12], un thésaurus spécifique à un domaine d'application est consulté afin de procéder à une recherche orientée selon le domaine particulier. Cependant, il est difficile d'obtenir un tel thésaurus qui assure une bonne couverture du vocabulaire et des relations utiles pour la RI. En effet, toutes les relations utiles pour la RI ne sont pas nécessairement justifiées linguistiquement. Par exemple, il serait difficile d'établir une relation formelle directe entre « mondialisation » et « aide humanitaire » dans un thésaurus. Cependant, cette relation est fortement utile pour les besoins de la RI.

Un autre problème avec les thésaurus est que les relations stockées sont généralement non pondérées. Même s'il est possible de déterminer des poids à ces relations selon les types ou selon la topologie des relations [21], il est difficile de distinguer précisément une relation forte et une relation faible. Par conséquent, on est souvent contraint à utiliser les relations avec une pondération sommaire voire uniforme, ce qui mène à des résultats discutables [28].

Pour la seconde approche – créer un modèle statistique automatiquement, il n'est pas nécessaire de recourir à des ressources construites manuellement pour représenter le domaine. Les caractéristiques du domaine sont extraites automatiquement par l'entremise d'un ensemble de documents classés dans le domaine en question. On peut voir plusieurs autres avantages : Les relations extraites sont pondérées selon leurs occurrences dans les documents ; les relations moins strictes mais utiles pour la RI (par exemple, la relation de cooccurrence) peuvent être également captées.

Dans cet article, nous adoptons la seconde approche. Afin d'établir un modèle statistique pour un domaine, nous supposons que chaque domaine possède un ensemble de documents exemples. Ces documents caractérisent le domaine de différentes façons : Ils spécifient un vocabulaire spécifique du domaine, et ils spécifient les relations lexicales (de cooccurrence) entre les termes. Les modèles construits seront exploités de trois façons différentes dans cette étude : pour étendre la requête, pour réordonner les documents retrouvés et pour lisser sémantiquement le modèle du document en utilisant les relations lexicales extraites. Nos expériences ont montrées de façon concluante que toutes ces exploitations du domaine aboutissent à des améliorations sensibles sur la performance de recherche. Ainsi, nous concluons que la prise en

compte du domaine d'intérêt est bénéfique, et que le modèle de langue statistique est un cadre approprié.

La suite de cet article est organisé comme suit : Nous allons d'abord discuter de la RI contextuelle dans Section 2. Les modèles de langue de base seront décrits dans Section 3. La construction d'un modèle de domaine sera présentée dans Section 4. Nous allons décrire trois façons d'exploiter les modèles de domaine dans Section 5. Nous décrirons les expérimentations dans Section 6 avant de tirer des conclusions dans Section 7.

2. RI CONTEXTUELLE

Bien que l'importance du contexte soit reconnue en RI [8], il existe peu de modèles opérationnels qui l'intègrent dans les différentes étapes de la recherche. La raison en est que nous ignorons encore à ce jour quelles sont les entités contextuelles devant être considérées, et surtout comment ils doivent être représentées et intégrées dans le processus de recherche d'information. En dépit de cette situation, on voit apparaître depuis quelques années en RI des approches qui considèrent certains aspects du contexte.

Ainsi, dans [15], la structure de similarité du corpus est considérée. Les documents similaires sont regroupés pour former des agrégats (clusters), représentant les thèmes ou domaines abordés dans la collection. Ces agrégats sont utilisés pour étendre la portée discursive des documents, par exemple, en combinant (lissant) le modèle du document par le modèle du domaine. Soulignons que la qualité de cette approche dépend largement de la capacité de la méthode d'agrégation à capter adéquatement la structure de similarité du corpus. Dans la présente étude, nous nous intéressons non pas au domaine du document, mais plutôt au domaine de la requête.

Considérant le contexte cognitif de l'usager, [14] exploite l'expertise de l'usager dans un cadre logique. La requête est désambiguïsée en tenant compte des croyances de l'usager inférées des documents consultés antérieurement. L'interprétation de la requête la plus compatible avec les croyances de l'usager est retenue. Bien que les résultats obtenus soient fort intéressants, en pratique, il est très difficile d'extraire les relations logiques strictes à partir des documents consultés. Ainsi, d'autres études utilisent plutôt une modélisation plus flexible du domaine.

Il est également possible de contextualiser une requête en exploitant des évidences implicites inférées du comportement de l'usager (par rétroaction de pertinence). Plusieurs facteurs peuvent refléter l'appréciation de l'usager [9], comme la consultation, la sauvegarde et le mouvement oculaire. Dans [23], on propose de redéfinir la requête d'après les documents consultés durant la session. Cependant, le résultat de cette redéfinition est considéré seulement pour la session courante, et il n'a pas un effet à plus long terme.

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 façons. Dans [6], une hiérarchie des domaines préconisés par l'usager est utilisée. [10] considère plutôt l'ensemble des documents consultés antérieurement par l'usager comme une spécification d'un domaine. Dans [26], 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 de l'usager.

Dans notre étude, nous considérons qu'un domaine d'intérêt peut être spécifié par un ensemble de documents. Un modèle de langue est construit pour chaque domaine. En spécifiant un domaine d'intérêt, l'usager peut aider à préciser sa requête.

Le domaine d'intérêt est intégré dans le processus de recherche dans une approche de modélisation de langue statistique. Cette approche est choisie parce qu'un modèle de langue statistique est capable de refléter les caractéristiques du domaine d'une façon flexible. Ce modèle a aussi une bonne tolérance au bruit qui peut se présenter dans les documents. Dans la prochaine section, nous allons décrire brièvement le principe de modélisation statistique de langue pour la RI.

3. MODELES DE LANGUE STATISTIQUE

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 $P(Q|D)$. Pour simplifier, on suppose en général que les mots qui apparaissent dans la requête sont indépendants. Ainsi, pour une requête $Q = t_1 t_2 \ldots t_n$, cette probabilité de génération est estimée comme suit :

$$P(Q \mid D) = P(t_1 t_2 ..t_n \mid D) = \prod_{t \in Q} P(t \mid D)^{c(t;Q)} \quad (1)$$

où $c(t;Q)$ est la fréquence du terme t dans la requête Q, et θ_D est le modèle du document, qui reflète la distribution de terme dans D. $P(t|\theta_D)$ est la probabilité du terme t dans le modèle du document.

Suivant le cadre de travail proposé dans [13], la similarité entre un document et une requête peut également être exprimée par la mesure de divergence de Kullback-Leibler (KL-divergence). La KL-divergence exprime la distance entre deux distributions probabilistes. D'un point de vue de la théorie de l'information, il s'agit de mesurer le coût supplémentaire nécessaire pour encoder la requête dans le modèle du document, ou bien l'entropie relative. La fonction d'ordonnancement est la suivante :

$$\begin{aligned} Score(Q,D) &= -KL(\theta_Q \parallel \theta_D) = \sum_{t \in V} P(t \mid \theta_Q) \log \frac{P(t \mid \theta_D)}{P(t \mid \theta_Q)} \\ &= \sum_{t \in V} P(t \mid \theta_Q) \log P(t \mid \theta_D) - \sum_{t \in V} P(t \mid \theta_Q) \log P(t \mid \theta_Q) \\ &\propto \sum_{t \in V} P(t \mid \theta_Q) \log P(t \mid \theta_D) \end{aligned} \quad (2)$$

où θ_Q est le modèle de la requête Q, et V est l'ensemble de vocabulaire dans la langue. Remarquons que la dernière simplification est faite parce que $\sum_{t \in V} P(t \mid \theta_Q) \log P(t \mid \theta_Q)$ ne dépend que de la requête, et n'affecte donc pas le rang des documents. Ainsi, cette fonction peut être vue comme une entropie croisée.

Soulignons qu'il existe une relation entre les équations (1) et (2) : Les deux mesures sont équivalentes du point de vue du rang des documents, si le modèle de langue θ_Q pour la requête est estimé par fréquence relative des mots-clefs dans la requête. En effet :

$$\begin{aligned} &-KL(\theta_Q \parallel \theta_D) = \sum_{t \in V} P(t \mid \theta_Q) \log \frac{P(t \mid \theta_D)}{P(t \mid \theta_Q)} \propto \sum_{t \in V} P(t \mid \theta_Q) \log P(t \mid \theta_D) \\ &\propto \sum_{t \in V} \frac{c(t;Q)}{|Q|} \log P(t \mid \theta_D) \propto \sum_{t \in V} c(t;Q) \log P(t \mid \theta_D) = \log P(Q \mid D) \end{aligned} \quad (3)$$

Une opération importante est le lissage de modèle, et ce en particulier pour le modèle de document. En effet, sans un lissage, si un terme *t* de la requête est absent d'un document, alors sa probabilité $P(t \mid \theta_D) = 0$, conduisant à $Score(Q, D) \to -\infty$. Or, un document dans lequel un terme de la requête est absent peut aussi être pertinent. Pour éviter ce problème, il s'avère essentiel de lisser le modèle du document par un modèle de retrait, qui est généralement le modèle de la collection θ_C [30]. Une des stratégies fréquemment utilisées est le lissage de Jelinek-Mercer, soit :

$$P(t \mid \theta'_D) = (1 - \lambda) P(t \mid \theta_D) + \lambda P(t \mid \theta_C) \quad (4)$$

Il est montré que ce lissage a aussi pour effet de modéliser la spécificité des termes de la requête, car $P(t \mid \theta'_D)$ incorpore un facteur similaire à IDF [30]. Notons que dans les travaux antérieurs, le modèle de requête θ_Q est généralement estimé par la fréquence relative sans lissage.

4. CONTEXTE USAGER

Comme nous avons mentionné, nous envisageons le scénario d'utilisation suivant pour l'identification d'un domaine particulier : pour chaque requête adressée au système, l'usager identifie un domaine pour son besoin d'information, soit l'une des catégories de la hiérarchie préétablie, par exemple « Finance », « 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. La spécification des domaines, qui correspondent aux intérêts à long terme de l'usager, peut aussi être faite via un profil d'usager.

Pour chaque domaine, nous associons un ensemble de documents du domaine. Nous supposons que ces documents caractérisent le domaine d'une façon implicite. Deux questions se posent : Comment constituer un ensemble de documents exemples dans chaque domaine, et comment exploiter ces documents ?

Pour constituer un ensemble de documents pour un domaine, nous pouvons envisager deux stratégies :

- Alimenter cet ensemble caractéristique du domaine au fur et à mesure par les documents que l'usager consulte au cours des multiples sessions. On suppose ici que ces documents consultés sont toujours

dans le domaine d'intérêt de l'usager, que ce domaine soit spécifié en même temps que la requête ou à travers un profil usager.
- Exploiter une classification manuelle ou automatique des documents du domaine et considérer les documents de la classe correspondante.

Dans nos travaux, nous avons testé ces deux approches. Pour la première approche, tous les documents que les usagers du système ont jugés pertinents pour des requêtes antérieures dans le même domaine sont collectés pour constituer des exemples du domaine. Cette approche a toutefois l'inconvénient de ne capter que certains aspects du domaine. En effet, la caractérisation du domaine dépend dans ce cas-ci des requêtes émises antérieurement. L'alternative est d'utiliser une hiérarchie de classes existantes, soit ODP [1] dans notre cas, pour identifier un ensemble de documents spécifiques à chacun des domaines. Cette approche est plus intéressante dans la mesure où la majorité des facettes du domaine sont captées.

Pour répondre à la seconde question - l'exploitation de cet ensemble de documents, il est important de souligner le fait que les documents ne caractérisent qu'une façon implicite et non stricte le domaine d'intérêt, par exemple le vocabulaire spécifique du domaine. Il existe aussi beaucoup d'autres éléments non informatifs - le bruit. Il est important de dégager les éléments significatifs par un moyen résistant au bruit. Les modèles de langues sont appropriés pour jouer ce rôle, car ils possèdent cette capacité de refléter les caractéristiques importantes dans un environnement bruité. Nous proposons donc d'établir un modèle de langue pour chaque domaine.

Il est possible de construire un modèle de langue directement à partir d'un ensemble de documents. Cependant, dû à l'existence du bruit, ce qui est extrait de cette façon n'est pas un modèle spécifique du domaine, mais un modèle mixte qui mélange le domaine spécifique et le domaine général. En effet, dans n'importe quel texte dans un domaine particulier, on utilise en alternance un vocabulaire général et un vocabulaire spécifique. Cette approche naïve ne permettrait pas de dégager suffisamment les éléments spécifiques du domaine.

Pour corriger cette situation, nous utilisons une procédure d'extraction du modèle spécifique comme suit : nous considérons que chaque document du domaine est généré conjointement par un modèle du domaine spécifique et un modèle plus général, correspondant à l'usage

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

courant de la langue générale. Ainsi, la probabilité de génération d'un document dans un domaine est exprimée comme suit :

$$P(D \mid \theta'_{Dom}) = \prod_{t \in D} [(1-\eta)P(t \mid \theta_{Dom}) + \eta P(t \mid \theta_C)]^{c(t;D)} \quad (5)$$

où θ_C est le modèle de la collection utilisé pour représenter le modèle général, θ_{Dom} est le modèle du domaine spécifique que l'on cherche à extraire, η est un paramètre de lissage que nous allons fixer, et θ'_{Dom} est le modèle mixte engendré par les deux sources – ce qui est directement construit à partir de l'ensemble de documents.

Afin d'estimer θ_{Dom}, nous utilisons un algorithme EM [7, 29], qui détermine θ_{Dom} de façon à maximiser $P(Dom \mid \theta'_{Dom})$, où Dom est l'ensemble des documents du domaine. Cette procédure d'extraction aboutit au meilleur modèle du domaine possible pour générer les documents exemples. Plus formellement :

$$\begin{aligned} \theta_{Dom} &= \arg\max_{\theta_{Dom}} P(Dom \mid \theta'_{Dom}) = \\ &= \arg\max_{\theta_{Dom}} \prod_{D \in Dom} \prod_{t \in D} [(1-\eta)P(t \mid \theta_{Dom}) + \eta P(t \mid \theta_C)]^{c(t;D)} \end{aligned} \quad (6)$$

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

$$w^{(i)}(t) = \frac{(1-\eta)P^{(i)}(t \mid \theta_{Dom})}{(1-\eta)P^{(i)}(t \mid \theta_{Dom}) + \eta P(t \mid \theta_C)} \quad \text{(E-Étape)} \quad (7)$$

$$P^{(i+1)}(t \mid \theta_{Dom}) = \frac{\sum_{D \in Dom} c(t;D) w^{(i)}(t)}{\sum_{t'} \sum_{D \in Dom} c(t';D) w^{(i)}(t')} \quad \text{(M-Étape)} \quad (8)$$

Essentiellement, ce processus itératif tente d'éliminer une partie (η - qui est un paramètre fixe) du modèle construit à partir de l'ensemble de documents, considérée correspondre à la langue générale. Au terme des itérations, il en résulte un modèle de langue dont la distribution de la masse de probabilité se concentre sur les termes spécifique du domaine, et peu observés dans le modèle général.

Le tableau I montre la variation des probabilités de certains termes observés dans le domaine « Environment » de TREC [2] au cours des

[2] Text Retrieval Conference: http://trec.nist.gov/

itérations de l'algorithme EM. Après 10 itérations, nous observons que les probabilités des termes spécifiques, tels que « pollution », « smog », « toxic » et « greenpeace », ont augmentées significativement, alors que les probabilités des termes courants, tels que « report », « time », « company » et « 1 », ont diminuées drastiquement.

Termes	Prob. initiale	Prob. résultante	Variation relative	Termes	Prob. initiale	Prob. résultante	Variation relative
pollution	0.0061	0.0102	+ 66%	million	0.0030	0.0016	- 46%

de la requête par le modèle du domaine. Il en résulte une requête étendue qui, par l'entremise des termes additionnels du domaine, exprime mieux le besoin d'information de l'usager. En effet, l'ajout de termes spécifiques du domaine peut accroître le pouvoir d'expression de la requête tout en la désambiguïsant, comme nous l'avons illustré plus tôt.

Intuitivement, l'approche adoptée se résume à une expansion de requête réalisée dans le cadre formel des modèles de langue. Mais cette expansion dépend du domaine. En effet, en interpolant le modèle du domaine au modèle original de la requête, on redistribue une partie de la masse de probabilité du modèle de la requête sur les termes caractéristique du domaine. Ces derniers sont les termes qui auraient pu être utilisés dans la requête par l'usager. Formellement, le nouveau modèle de la requête peut être obtenu comme suit :

$$\theta'_Q = (1 - \alpha)\theta_Q + \alpha\theta_{Dom} \quad (9)$$

où θ_Q est le modèle original de la requête qui est estimé directement par la fréquence relative des termes dans Q, θ_{Dom} est le modèle du domaine, et θ'_Q est le modèle engendré par les deux précédents. α est un paramètre de lissage.

Bien entendu, θ_{Dom} n'est pas le seul modèle pouvant être combiné à θ_Q. Un autre modèle fort utile est celui construit sur une rétroaction de pertinence, c'est-à-dire à partir des documents qui se retrouvent en tête de la liste de résultats. Ainsi, en utilisant les n premiers documents retrouvés par la requête initiale, nous estimons un modèle additionnel θ_R. Ce modèle peut être combiné aux précédents (via un autre paramètre de lissage β), donnant le modèle suivant pour la requête :

$$\theta'_Q = (1 - \alpha - \beta)\theta_Q + \alpha\theta_{Dom} + \beta\theta_R \quad (10)$$

Tout comme θ_{Dom}, θ_R est estimé en appliquant l'algorithme EM décrit à la section 4 sur les documents considérés. L'intérêt du modèle mixte défini en (10) est de considérer deux sources d'informations distinctes, soient θ_{Dom}, qui modélise le discours du domaine, et θ_R, qui reflète ce que l'usager recherche. L'équation (10) est donc un modèle plus complet de la requête.

En pratique, un domaine peut être plus ou moins large. Pour un domaine large, les documents exemples ne sont pas concentrés sur certains thèmes, et le vocabulaire peut disperser. Un tel modèle de domaine complet pourrait introduire du bruit dans le modèle de requête lorsqu'il est interpolé avec le modèle de la requête. Pour résoudre ce problème, il est également possible de créer un sous domaine selon la

requête, en considérant seulement les *n* premiers documents du domaine jugés les plus pertinents pour la requête. Cette approche est justifiée par le fait que l'opération d'expansion de requête est une opération critique. Les termes qui y sont introduits doivent être pertinents pour la requête courante, sans quoi un glissement thématique de la requête se produit. Pour un domaine aussi vague que « Science et Technologie » par exemple, les documents peuvent porter sur des thèmes très différents. Ainsi, il est avantageux de limiter le discours du domaine aux documents qui s'apparentent le plus à la requête.

Une fois le nouveau modèle de la requête θ'_Q obtenu, nous utilisons une KL-divergence négative pour déterminer le rang de chaque document de la collection. Le pointage d'un document *D* pour une requête *Q* du domaine *Dom* est déterminé comme suit :

$$\begin{aligned} Score(Q, D, Dom) &= \sum_{t \in V} P(t \mid \theta'_Q) \log \frac{P(t \mid \theta'_D)}{P(t \mid \theta'_Q)} \\ &\propto \sum_{t \in V} P(t \mid \theta'_Q) \log P(t \mid \theta'_D) \end{aligned} \tag{11}$$

où θ'_D est le modèle lissé du document tel que défini en (4).

5.2. Réordonner les documents retrouvés

Une seconde stratégie est d'utiliser le modèle du domaine pour réordonner les documents retrouvés, comme dans [6, 10, 26]. Les documents sont d'abord ordonnés selon leur relation à la requête. Ensuite, l'ordre des documents est modifié d'après leur correspondance avec le domaine. Cette approche permet de favoriser les documents qui s'apparentent les plus au domaine de la requête.

La fonction d'ordonnancement proposée est une combinaison des KL-divergences négatives, reflétant respectivement la correspondance du document à la requête *Q* et au domaine *Dom*. Nous utilisons la fonction suivante pour calculer un nouveau score :

$$Score(Q, D, Dom) = -\left[(1-\gamma) \underbrace{KL(\theta_Q \parallel \theta'_D)}_{Similarité\ Q-D} + \gamma \underbrace{KL(\theta_D \parallel \theta'_{Dom})}_{\text{Similarité D-Dom}} \right] \tag{12}$$

Dans cette fonction, nous avons utilisé un coefficient γ pour contrôler l'importance relative accordée au domaine. Lorsque γ = 0, le modèle

redevient l'approche de base, qui ne considère que la requête et le document.

5.3. Exploiter les relations lexicales du domaine

Dans les deux approches précédentes, nous avons exploité la distribution de termes dans les documents du domaine. Mis à part une distribution particulière, les documents du domaine peuvent aussi nous révéler les relations possibles entre les termes. Par exemple, dans le domaine « Finance », le terme « budget » est fortement relié au terme « planification ». Cette relation peut être utilisée pour effectuer une expansion de la requête sur « planification ». Ainsi, dans ce troisième modèle, on exploite les dépendances lexicales du domaine.

Comme nous avons mentionné, il est possible d'exploiter un thésaurus pour obtenir des relations lexicales. Dans notre étude, ces relations sont extraites selon les cooccurrences. Inspiré des travaux [3] et [5], 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 suivant les dépendances lexicales entre les termes dans le domaine. Durant le lissage, un terme fortement lié se voit attribuer une probabilité plus grande qu'un autre terme non relié. Ainsi, nous pouvons parler d'un lissage sémantique.

Concrètement, étant donné une requête Q et le domaine Dom, la probabilité d'observer le document D peut être exprimée comme suit :

$$P(D \mid Q, Dom) = \frac{P(Q \mid D, Dom)P(D \mid Dom)}{P(Q \mid Dom)} \propto P(Q \mid D, Dom)P(D \mid Dom) \qquad (13)$$

La probabilité $P(D|Dom)$ peut être estimée comme une probabilité de génération, soit :

$$P(D \mid Dom) = \prod_{t \in D} P(t \mid \theta'_{Dom})^{c(t;D)} \qquad (14)$$

La probabilité $P(Q|D,Dom)$ traduit la probabilité de générer la requête étant donné le document et les dépendances lexicales inférées du domaine. Le modèle combinant D et Dom est dénoté par le modèle de dépendance $\Phi_{D,Dom}$. La probabilité de chaque terme t de Q dans ce modèle est estimé en considérant sa dépendance $t_{Dom}(t \mid d)$ à chaque terme d dans D. Ainsi :

$$P(t \mid \Phi_{D,Dom}) = \sum_{d \in D} t_{Dom}(t \mid d) P(d \mid \theta'_D) \qquad (15)$$

où θ'_D est un modèle lissé du document et $t_{Dom}(t|d)$ est la probabilité de dépendance de t à d estimée dans *Dom*. Cette probabilité de dépendance est basée sur la cooccurrence des termes dans les documents du domaine. Dans notre étude, nous considérons une fenêtre de 5 mots pour la cooccurrence. Soit $c(<t,d>;D)$, le nombre de fois que t co-occurre avec d dans un document D du domaine *Dom*. Nous définissons la dépendance lexicale comme suit :

$$t_{Dom}(t \mid d) = \frac{\sum_{D \in Dom} c(\langle t,d \rangle; D)}{\sum_{t' \in V} \sum_{D \in Dom} c(\langle t'.d \rangle; D)} \tag{16}$$

Remarquons que les documents du domaine peuvent couvrir seulement une partie de relations lexicales utiles. Une autre partie de relations utiles peut se trouver dans la langue générale. Ainsi, nous devons aussi considérer les dépendances lexicales dans cette dernière, reflétée par toute la collection. Le modèle de dépendance est donc redéfini comme suit :

$$P(t \mid \Phi'_{D,Dom}) = \sum_{d \in D} [(1-\mu) t_{Dom}(t \mid d) + \mu t_c(t \mid d)] P(d \mid \theta'_D) \tag{17}$$

Finalement, en plus de l'expansion utilisant les relations lexicales, le modèle $\Phi'_{D,Dom}$ peut aussi être lissé par le modèle uni-gramme du document θ'_D. Ainsi, le modèle final utilisé est le suivant :

$$P(t \mid \Phi''_{D,Dom}) = (1-\lambda) P(t \mid \Phi'_{D,Dom}) + \lambda P(t \mid \theta'_D) \tag{18}$$

où λ est un paramètre de lissage. Par conséquent, $P(Q|D)$ est estimé comme suit :

$$P(Q \mid D, Dom) = \prod_t P(t \mid \Phi''_{D,Dom})^{c(t;Q)} \tag{19}$$

Substituant les équations (14) et (19) dans l'expression logarithmique de (13), on obtient :

$$\begin{aligned} \log P(D \mid Q, Dom) &= \sum_{t \in Q} c(t;Q) \log P(t \mid \Phi''_{D,Dom}) + \sum_{t \in D} c(t;D) \log P(t \mid \theta'_{Dom}) \\ &\propto |Q| \sum_{t \in Q} P(t \mid \theta_Q) \log P(t \mid \Phi''_{D,Dom}) + |D| \sum_{t \in D} P(t \mid \theta_D) \log P(t \mid \theta'_{Dom}) \end{aligned} \tag{20}$$

Les deux composantes de l'équation (20) 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 $(1-\delta)$ et δ. Ainsi :

$$Score(Q, D, Dom) = -\left[(1-\delta)KL(\theta_Q \parallel \Phi''_{D,Dom}) + \delta KL(\theta_D \parallel \theta'_{Dom})\right] \quad (21)$$

Finalement, soulignons que lorsque λ = 1, le modèle de dépendance est ignoré. Le modèle défini en (21) se réduit alors au modèle de ré-ordonnancement défini en (12).

6. EXPERIMENTATIONS ET RESULTATS

Nos expériences ont été réalisées en utilisant les outils de Lemur [3]. Deux collections de TREC ont été utilisées, soient les documents des disques 1-2, et les documents 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 de la requête. Ainsi, il est possible de simuler nos approches. Les requêtes sont distribuées sur 13 domaines, tels que « Environment », « Finance », « Military », « Science and Technologies », etc. Pour les traitements de base, nous utilisons l'algorithme de Krovetz [11] pour la lemmatisation, et les termes fonctionnels sont retirés.

Suivant les deux approches mentionnées à la section 4, les documents pour chaque domaine sont récupérés des collections de TREC et de l'annuaire web d'ODP. Dans le premier cas, nous incluons les documents jugés pertinents pour les requêtes 51 à 150 dans leurs domaines respectifs. 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 inclus dans le domaine. Dans le cas des documents issus des répertoires d'ODP, les domaines d'ODP n'étant pas identiques à ceux des requêtes TREC, nous devons établir une correspondance. Idéalement, cette correspondance doit être établie automatiquement. Mais dans cette étude, nous l'avons fait manuellement pour simuler ce processus : nous avons créé une correspondance entre les domaines qui nous semblent s'apparenter. Par exemple, pour le domaine « Environment », les catégories suivantes ont été retenues : « Science/Environment », « Science/ Biology/Ecology » et « Society/Issues/Environment ». Remarquons que nous avons établi cette correspondance selon notre intuition sans chercher à favoriser notre approche.

Pour nos évaluations, les mesures de performance suivantes sont utilisées : la précision moyenne non-interpolée (préc.), le rappel pour les

[3]. The Lemur Toolkit: http://www.lemurproject.org

1000 premiers documents retrouvés (rappel), et la précision moyenne pour les 5 premiers documents retrouvés (préc. top 5). Le but de nos expériences est d'évaluer la contribution du domaine à différents niveaux du processus de recherche d'information. Les approches utilisant les informations du domaine seront comparées à une méthode de base qui n'exploite aucune information contextuelle. L'objectif 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 d'intérêt dans la recherche et que les méthodes que nous avons proposées sont raisonnables.

Chacun des paramètres utilisés dans nos modèles (notamment, les paramètres de lissages) ont été déterminés en explorant un espace des valeurs discrètes. Dans la mesure où leurs valeurs optimales ne semblent pas dépendre du jeu de donnée, nous avons pu fixer les paramètres sans risque de sur-apprentissage. Dans nos expérimentations, 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 lissage du modèle de la requête α, β et λ, et les coefficients d'interpolations de similarité γ et δ. Ainsi, dans la pratique, seulement quelques paramètres doivent êtres évalués avec soin.

6.1. Compléter la requête

Afin d'évaluer la contribution du modèle du domaine au modèle de la requête, nous avons comparé les performances du modèle étendu de la requête selon le domaine (les équations (9) et (11)) aux performances d'un modèle de base (l'équation (2)). Le modèle de base est un cas particulier du modèle étendu selon le domaine (lorsque $\alpha = 0$). Comme il a été suggéré dans la section 5.1, le modèle du domaine est estimé d'après les 20 premiers documents du domaine jugés les plus pertinents pour la requête en cours selon l'équation (2).

Dans le tableau II, nous pouvons observer des améliorations de 13% à 19% sur la précision moyenne lorsque le modèle du domaine est interpolé au modèle original de la requête. Cette comparaison montre clairement que le modèle du domaine permet de compléter l'information partielle de la requête. La requête ainsi complétée spécifie mieux le besoin d'information de l'usager.

Collection	Mesure	Modèle de base	Modèle étendu selon le domaine			
			TREC		ODP	
Disque 1-2	préc.	0.1867	0.2115	+13.28%	0.2205	+18.10%
	rappel	0.4384	0.4803	+9.55%	0.5008	+14.23%
	préc. top 5	0.4040	0.4420	+9.40%	0.4700	+16.33%
Disque 3	préc.	0.1770	0.2112	+19.32%	0.2086	+17.85%
	rappel	0.4270	0.4811	+12.66%	0.4946	+15.83%
	préc. top 5	0.3720	0.4480	+20.43%	0.4300	+15.59%

TAB. II – *Performances du modèle étendu de la requête selon le domaine.*

Collection	Mesure	Modèle étendu par rétroaction de pertinence	Modèle étendu mixte			
			TREC		ODP	
Disque 1-2	préc.	0.2700	0.2899	+7.37%	0.2901	+7.44%
	rappel	0.5393	0.5704	+5.76%	0.5736	+6.36%
	préc. top 5	0.5220	0.5200	-0.38%	0.5160	-1.14%
Disque 3	préc.	0.2462	0.2604	+5.76%	0.2599	+5.56%
	rappel	0.5242	0.5437	+3.71%	0.5449	+3.94%
	préc. top 5	0.4740	0.4720	-0.42%	0.4680	-1.26%

TAB. III – *Performances du modèle étendu mixte de la requête.*

Dans le but de vérifier s'il demeure intéressant de considérer le domaine d'intérêt de l'usager lorsque d'autres sources d'informations sont considérées, nous avons également comparé notre approche avec un modèle étendu par rétroaction de pertinence : On considère conjointement le modèle du domaine et le modèle de rétroaction de pertinence (voir les équations (10) et (11)). Le modèle ainsi étendu est un cas particulier du second modèle ($\alpha = 0$). Tout comme pour le modèle du domaine, seuls les 20 premiers documents retrouvés pour la requête en cours sont considérés pour estimer le modèle de rétroaction de pertinence.

Le tableau III décrit les performances de chacun des modèles. Malgré le fait que le modèle de rétroaction soit plus apte à préciser le besoin de l'usager que le modèle du domaine, il est intéressant de constater que l'ajout du modèle du domaine demeure bénéfique même quand la rétroaction de pertinence est utilisée : La précision moyenne est encore améliorée de 6%-7%. Ce résultat suggère que le modèle du domaine exprime des informations que le modèle de rétroaction de pertinence n'exprime pas. Nous expliquons ceci par le fait que les documents de la rétroaction de pertinence sont parfois trop liés à la requête, empêchant la requête d'être pleinement étendue. En ajoutant le domaine, on permet une extension supplémentaire. Cette extension supplémentaire a surtout un

effet sur le rappel. Ceci est confirmé par les deux faits suivants : pour les premiers documents retrouvés, on observe une légère baisse de la précision (préc. top 5). Mais globalement, la précision moyenne et le rappel sont augmentés.

Il est également intéressant d'observer que le modèle du domaine apporte la même contribution qu'il soit défini avec les documents pertinents des autres requêtes du domaine ou avec les documents d'ODP. En regard aux performances mesurées, les deux modèles sont similaires. Ceci montre que la méthode pour estimer le modèle du domaine résiste bien au bruit.

Collection	Mesure	Modèle de base	Modèle de ré-ordonnancement selon le domaine			
			TREC		ODP	
Disque 1-2	préc.	0.1867	0.2134	+14.30%	0.2107	+12.85%
	rappel	0.4384	0.4750	+8.34%	0.4713	+7.50%
	préc. top 5	0.4040	0.4620	+14.35%	0.4440	+9.90%
Disque 3	préc.	0.1770	0.1922	+8.58%	0.1893	+6.94%
	rappel	0.4270	0.4505	+5.50%	0.4491	+5.17%
	préc. top 5	0.3720	0.4260	+14.51%	0.4360	+17.20%

TAB. IV – *Performances du modèle de ré-ordonnancement.*

6.2. Réordonner les documents

Dans la seconde expérimentation, nous voulions mesurer les avantages qu'il y a à réordonner les 3000 premiers documents retrouvés selon leur degré d'appartenance au domaine. Les performances d'un modèle de base ont été comparées à celles du modèle de ré-ordonnancement (voir l'équation (12)). Le modèle de base est un cas particulier du modèle de ré-ordonnancement ($\gamma = 0$).

Dans le tableau IV, nous constatons qu'il est avantageux de réordonner les documents de cette façon. Bien que les gains soient moindres que les gains obtenus avec l'approche précédente, il y une amélioration de 7% à 14% de la précision moyenne. De plus, en considérant le degré d'appartenance de chacun des documents retrouvés par rapport au domaine, la précision moyenne pour les 5 premiers documents est accrue de 10% à 17% selon les cas. Ceci montre que le ré-ordonnancement des résultats est aussi une méthode raisonnable d'exploiter le domaine, surtout pour augmenter la précision des premiers documents. Cette approche est particulièrement intéressant dans la situation où il n'est pas possible de modifier la fonction de recherche d'un système de RI existant (comme un

engin de recherche sur le Web). Nous pouvons alors utiliser une fonction supplémentaire sur la correspondance des documents retrouvés au domaine en question, pour les re-ordonner.

6.3. Exploiter les relations lexicales du domaine

La dernière série d'expérimentation porte sur le modèle défini à la section 5.3. Nous avons comparé les performances d'un modèle de base à celui du modèle de dépendance (voir l'équation (21)). Le modèle de base correspond à la configuration suivante : $\lambda = 1$ et $\delta = 0$.

Collection	Mesure	Modèle de base	Modèle de dépendance du domaine			
			TREC		ODP	
Disque 1-2	préc.	0.1867	0.2121	+13.61%	0.2101	+12.59%
	rappel	0.4384	0.4738	+8.07%	0.4714	+7.53%
	préc. top 5	0.4040	0.4620	+14.35%	0.4400	+8.91%
Disque 3	préc.	0.1770	0.1902	+7.46%	0.1888	+6.67%
	rappel	0.4270	0.4501	+5.41%	0.4498	+5.34%
	préc. top 5	0.3720	0.4260	+14.51%	0.4340	+16.67%

TAB. V – *Performances du modèle de dépendance du domaine.*

Dans le tableau V, nous pouvons observer un gain relatif de la précision moyenne lorsque le modèle de dépendance du domaine est utilisé. Les relations lexicales du domaine $t_{Dom}(t|d)$ sont considérées dans une proportion de 10% ($\mu = 0.9$). Le modèle de dépendance $\Phi'_{D,Dom}$ compte également pour 10% ($\lambda = 0.9$).

Durant une exploration partielle de l'espace des paramètres, il s'avère toutefois plus profitable de considérer seulement les relations lexicales extraites de la collection ($\mu = 1$). La raison en est que les relations lexicales couvertes par les relations spécifiques au domaine sont trop limitées. La couverture n'est pas suffisante. De plus, soulignons que les relations du domaine ne sont applicables que si le document appartient au domaine. Or, étant donné la composition de la fonction d'ordonnancement défini à l'équation (21), il s'avère difficile de calibrer efficacement le degré de confiance avec lequel les relations du domaine sont appliquées. L'importance accordée au domaine ne peut être accrue sans par le fait même atténuer l'importance accordée à la requête (voir δ).

En ce qui concerne les relations lexicales, il s'avère plus avantageux de considérer seulement le modèle uni-gramme. En effet, les performances sont plus élevées lorsque $\lambda = 1$. Notez que quand $\lambda = 1$, le modèle de

dépendance du domaine défini à la section 5.3 se résume au modèle de réordonnancement défini à la section 5.2. En comparant les résultats du tableau IV avec ceux du tableau V, nous constatons que la précision moyenne et le rappel sont légèrement supérieurs lorsque le modèle de réordonnancement est utilisé. La précision moyenne pour les premiers documents demeure toutefois la même. Il semble donc que les gains réalisés par le modèle de la section 5.3 soient en partie dus au fait que ce modèle prend en considération le degré d'appartenance du document au domaine. En pratique, le modèle de dépendance ne parvient pas à accroître la portée lexicale du document. La raison en est que les termes mis en relations dans le modèle de dépendances sont généralement déjà contenus dans le document. Il n'y a donc pas de véritable expansion du document en exploitant les relations lexicales.

Finalement, à travers nos expérimentations, nous pouvons voir que les trois façons d'intégrer le modèle du domaine aboutissent à une augmentation sensible des performances. Certes, les dépendances lexicales du domaine ne semblent pas avoir une importance primordiale dans l'évaluation de la pertinence d'un document, mais le modèle de dépendance offre tout de même des améliorations appréciables par rapport au modèle de base. Ces séries d'expérimentations montrent donc l'importance de tenir compte du domaine. Elles montrent aussi 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 de calcul en ligne pour l'évaluation d'une requête demeure court. Le modèle de dépendance nécessite toutefois beaucoup d'espace mémoire pour stocker les probabilités de dépendance entre les termes (1 Go), tandis que les autres approches nécessitent beaucoup moins d'espace.

7. CONCLUSION

Dans cet article, nous avons présenté trois modèles opérationnels pour exploiter un aspect du contexte cognitif de l'usager - son domaine d'intérêt. Caractérisé 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 afin d'étendre la requête.

Nos expériences avec ces trois approches 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 fortes améliorations de la performance dans les trois cas. L'approche la plus performante est celle qui utilise le domaine pour compléter la requête (section 5.1). Cela suggère que la plus grande faiblesse de la requête – l'incomplétude de spécification du besoin due à sa taille réduite, peut être compensée par la prise en compte du domaine. Le domaine d'intérêt de l'usager est manifestement 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 un modèle de pseudo-rétroaction de pertinence, par exemple. Cela a d'ailleurs produit une amélioration supplémentaire. Cependant, nous avons observé que même dans le cas où la pseudo-rétroaction de pertinence est utilisée, le modèle du domaine peut apporter une contribution supplémentaire.

Cette étude est encore préliminaire – notre objectif est de montrer l'utilité du domaine et le potentiel d'une approche qui tient compte du domaine. Plusieurs aspects peuvent être améliorés. Par exemple, nous avons considéré tous les domaines pour une requête 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 « Military » s'avère plus homogène que le domaine « Science and Technologies ». 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 peut être déterminé au moyen d'une classification automatique tout en offrant les mêmes performances.

8. REFERENCES

[1] Allan, J., Aslam, J., Belkin, N., Buckley, C., Callan, J., Croft, B., Dumais, S., Fuhr, N., Harman, D., Harper, D.J., Hiemstra, D., Hofmann, T., Hovy, E., Kraaij, W., Lafferty, J., Lavrenko, V., Lewis, D., Liddy, L, Mammatha, R., McCallum, A., Ponte, J., Prager, J., Radev, D., Resnik, P., Robertson, S., Rosenfield, R., Roukos, S., Sanderson, M., Schwartz, R., Singhal, A., Smeaton, A., Turtle, H., Voorhees, E., Weischedel, R., Xu, J., Zhai, C.,

Challenges in information retrieval and language modeling: Report of a workshop held at the center for intelligent information retrieval, *SIGIR Forum*, vol 37, pages 31-47, New-York : ACM, 2003.

[2] Belkin, N.J., Interaction with texts: Information retrieval as information seeking behavior, *Information Retrieval'93 : Von der modellierung zu anwendung*, pages 55-66, Konstanz : Krause & Womser-Hacker, 1993.

[3] Berger, A., Lafferty, J., Information retrieval as statistical translation, *SIGIR'99 : Proceedings of 22nd ACM International conference on research and development in information retrieval*, pages 222-229, New-York : ACM, 1999.

[4] Braslavski, P., Alshanski, G., Shishkin, A., ProThes: Thesaurus-based Meta-Search Engine for a Specific Application Domain, *WWW`04*, pages 222-223, New-York : ACM, 2004.

[5] Cao, G., Nie, J.-Y., Bai, J., Integrating word relationships into language models, SIGIR'05 : *Proceedings of 28th ACM International conference on research and development in information retrieval,* pages 298-305, New-York : ACM, 2005.

[6] Chirita, P.A., Paiu, R., Nejdl, W., Kohlschütter, C., Using ODP metadata to personalize search, SIGIR'05 : *Proceedings of 28th ACM International conference on research and development in information retrieval*, pages 178-185, New-York : ACM, 2005.

[7] Hiemstra, D., Robertson, S., Zaragoza, H., Parsimonious language models for information retrieval, SIGIR'04 : *Proceedings of 27th ACM International conference on research and development in information retrieval*, pages 178-185, New-York : ACM, 2004.

[8] Ingwersen, P., Jäverlin, K., Information retrieval in context: IRiX, *SIGIR Forum*, vol 39, pages 31-39, New-York : ACM, 2004.

[9] Kelly, D., Teevan, J., Implicit feedback for inferring user preference: A bibliography, *SIGIR Forum*, vol 37, pages 18-28, New-York : ACM, 2003.

[10] Kim, H.-R., Chan, P.K., Personalized ranking of search results with learned user interest hierarchies from bookmarks, *WEBKDD'05 Workshop* at the *11th ACM International conference on Knowledge discovery and data mining*, pages 32-43, New-York : ACM, 2005.

[11] Krovetz, R., Viewing morphology as an inference process, *SIGIR '93: Proceedings of the 16th ACM International conference on research and development in information retrieval*, pages 191-202, New-York : ACM, 1993.

[12] Kwon O.-W., Kim, M.-C., Choi K.-S., Query expansion using domain-adapted, weighted thesaurus in an extended Boolean model, CIKM '94: *Proceedings of the third international conference on Information and knowledge management,* pages 140-146, New-York : ACM, 1994.

[13] Lafferty, J., Zhai, C., Language models, query models, and risk minimization for information retrieval, SIGIR'01, *Proceedings of 24th ACM International conference on research and development in information retrieval*, pages 111-119, New-York : ACM, 2001.

[14] Lau, R., Bruza, P., Song, D., Belief revision for adaptive information retrieval, SIGIR'04, *Proceedings of 27th ACM International conference on research and development in information retrieval,* pages 130-137, New-York : ACM, 2004.

[15] Liu, X., Croft, B.W., Cluster-based retrieval using language models, SIGIR'04, *Proceedings of 24th ACM International conference on research and development in information retrieval,* pages 186-193, New-York : ACM, 2004.

[16] Medin, D.L., Schaffer, M.M., Context theory of classification learning, *Psychological review*, vol 85, pages 207-238, 1978.

[17] Mishler, E.G., Meaning in context: Is there any other kind?, *Harvard education review*, vol 49, pages 1-19, Harvard : Harvard University, 1979.

[18] Morris, R.C., Toward a user-centered information service, *Journal of the American Society for Information Science*, vol 45, pages 20-30, Mississauga : Wiley, 1994.

[19] 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, pages 135-141, Mississauga : Wiley, 1994.

[20] Pitkow, J., Schütze, H., Cass, T., Cooley, R., Turnbull, D., Edmonds, A., Adar, E., Breuel, T., Personalized Search, *Communications of ACM*, vol 45, pages 50-55, New-York : ACM, 2002.

[21] Resnik, P., Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language, *Journal of Artificial Intelligence Research,* Vol. 11, pages 95-130, 1999.

[22] Saracevic, T., Information science, *Journal of the American society for information science*, vol 50, pages 1051-1063, Mississauga : Wiley, 1999.

[23] Shen, X., Tan, B., Zhai, C., Context-sensitive information retrieval using implicit feedback, *SIGIR'05 : Proceedings of 28th ACM International conference on research and development in information retrieval,* pages 43-50, New-York : ACM, 2005.

[24] Spink, A., Losee, R.,M., Feedback in information retrieval, *Annual review of information science and technology*, Vol. 31, pages 31-78, Medford : Information Today, 1996.

[25] Taylor, R.S., Information use environments, *Progress in communication science*, vol 10, pages 217-255, Norwood: Ablex Publishing Corp, 1991.

[26] Teevan, J., Dumais, S.T., Horvitz, E., Personalizing search via automated analysis of interests and activities, *SIGIR'05 : Proceedings of 28th ACM*

Deleted: SIGIR'01

Deleted: 2001

International conference on research and development in information retrieval, pages 449-456, New-York : ACM, 2005.

[27] Vakkari, P., Savolainen, R., Dervin, B., Information seeking in context, pages 467, London : Taylor Graham, 1997.

[28] Voorhees, E.M., Query Expansion using Lexical-Semantic Relations, *SIGIR'94 : Proceedings of the 17th ACM International conference on research and development in information retrieval*, pages 61-69, New-York : ACM, 1994.

[29] Zhai, C., Lafferty, J., Model-based feedback in the language modeling approach to information retrieval, *CIKM'01 : Proceedings of 10th International conference on information and knowledge management*, pages 403-410, New-York : ACM, 2001.

[30] Zhai, C., Lafferty, J., A study of smoothing methods for language models applied to ad hoc information retrieval, *SIGIR'01 : Proceedings of 24th ACM International conference on research and development in information retrieval*, pages 334-342, New-York : ACM, 2001.