Expansion sélective de requêtes par apprentissage

Adrian-Gabriel Chifu* — **Josiane Mothe****

* *IRIT UMR5505, CNRS, Université de Toulouse, Université Paul Sabatier (France)*
** *IRIT UMR5505, CNRS, Université de Toulouse, ESPE (France)*
* *adrian.chifu@irit.fr* ** *josiane.mothe@irit.fr*

RÉSUMÉ. *Si l'expansion de requête automatique améliore en moyenne la qualité de recherche, elle peut la dégrader pour certaines requêtes. Ainsi, certains travaux s'intéressent à développer des approches sélectives qui choisissent la fonction de recherche ou d'expansion en fonction des requêtes. La plupart des approches sélectives utilisent un processus d'apprentissage sur des caractéristiques de requêtes passées et sur les performances obtenues. Cet article présente une nouvelle méthode d'expansion sélective qui se base sur des prédicteurs de difficulté des requêtes, prédicteurs linguistiques et statistiques. Le modèle de décision est appris par un SVM. Nous montrons l'efficacité de la méthode sur des collections TREC standards. Les modèles appris ont classé les requêtes de test avec plus de* 90% *d'exactitude. Par ailleurs, la MAP est améliorée de plus de* 11%*, comparée à des méthodes non sélectives.*

ABSTRACT. *Query expansion (QE) improves the retrieval quality in average, even though it can dramatically decrease performance for certain queries. This observation drives the trend to suggest selective approaches that aim at choosing the best function to apply for each query. Most of selective approaches use a learning process on past query features and results. This paper presents a new selective QE method that relies on query difficulty predictors. The method combines statistically and linguistically based predictors. The QE method is learned by a SVM. We demonstrate the efficiency of the proposed method on a number of standard TREC benchmarks. The supervised learning models have performed the query classification with more than 90% accuracy on the test collection. Our approach improves MAP by more than 11%, compared to the non selective methods.*

MOTS-CLÉS : *Recherche sélective d'information, Prédicteurs de difficulté, Difficulté de requête, Expansion de requête, Apprentissage.*

KEYWORDS: *Selective information retrieval, Difficulty predictors, Query expansion, Machine learning.*

CORIA 2014, pp. 257–272, Nancy, 18-21 mars 2014

1. Introduction

Le terme de *recherche d'information sélective* (RIS) est parfois utilisé pour désigner le traitement d'ensembles de documents de façon parallèle, avec des index disjoints (Arguello *et al.*, 2009). Elle peut également faire référence à la sélection de la configuration du moteur de recherche la plus appropriée selon le contexte et la requête. Ce défi trouve probablement ses origines dans la variabilité des requêtes et des systèmes (Buckley, 2004). En effet, aux jours d'aujourd'hui, il est largement admis que les résultats en termes de performance varient en fonction des requêtes et en fonction des systèmes ; un système pouvant être très performant sur certaines requêtes et au contraire obtenir de mauvais résultats sur d'autres requêtes alors qu'un autre système aura des résultats opposés. Il est donc naturel de penser qu'en adaptant le processus de recherche à chaque requête il est possible d'améliorer les résultats globaux de recherche. La difficulté de requête est une autre caractéristique qu'il est possible de considérer dans les systèmes (De Loupy et Bellot, 2000 ; Carpineto *et al.*, 2001). En effet, dans le cas de requêtes faciles, on peut considérer que n'importe quelle configuration raisonnable de système peut être utilisée puisque toutes obtiendront de bons résultats. Les choses sont différentes en considérant des requêtes dites difficiles, celles pour lesquelles la plupart des systèmes ne retrouvent pas ou peu de documents pertinents. Intuitivement, il semble possible de maximiser le gain en termes de documents pertinents retrouvés en utilisant des techniques appropriées.

La sélection de la meilleure configuration en fonction des requêtes peut améliorer les résultats. Par exemple, Bigot *et al.* ont montré qu'il était possible d'apprendre la meilleure configuration de système en utilisant un échantillon de documents et des jugements de pertinence (Bigot *et al.*, 2011). De la même façon, Peng *et al.* ont proposé de choisir la meilleure fonction d'ordonnancement (*Learning to rank*) en fonction des requêtes (Peng *et al.*, 2010). Santos *et al.* de leur côté présentent une méthode selective de diversification des résultats (Santos *et al.*, 2010).

D'autres approches se concentrent sur le cas des requêtes difficiles et sur l'expansion sélective de requête. Une des hypothèses principales est que l'expansion de requête et spécifiquement l'expansion basée sur le *Pseudo Retour de Pertinence* (PRF) peut détériorer les résultats pour les requêtes avec des performances initiales faibles puisque l'expansion sera basée sur des documents non pertinents. Plusieurs auteurs proposent d'appliquer sélectivement l'expansion en se basant sur des prédicteurs de difficulté des requêtes. L'idée est d'éviter l'application de l'expansion quand elle va probablement diminuer la performance (Amati *et al.*, 2004 ; Cronen-Townsend *et al.*, 2004 ; Chen *et al.*, 2012). Au lieu de simplement décider si l'expansion doit être appliquée ou non, d'autres travaux essayent d'appliquer l'expansion différemment selon les requêtes (He et Ounis, 2004 ; Cao *et al.*, 2008 ; Lv et Zhai, 2009). Beaucoup de ces approches s'appuient sur la capacité à distinguer les requêtes difficiles. Ainsi, la prédiction de la difficulté des requêtes est un sujet majeur (Carmel et Yom-Tov, 2010) et des approches actuelles ont tendance à combiner diverses caractéristiques pour prédire la difficulté, généralement basées sur des statistiques sur les requêtes.

Dans ce papier, nous considérons l'expansion de requête comme un problème clé de la recherche d'information sélective. Il faut noter qu'il existe de nombreuses façons de réaliser l'expansion de requête. Dans son article de synthèse, Carpineto distingue l'expansion par retour de pertinence, le raffinement interactif des requêtes, la désambiguïsation des termes et la classification basée sur les résultats de recherche (Carpineto et Romano, 2012). Dans le travail présenté ici, nous nous appuyons sur les deux premiers types d'expansion.

Les approches de la littérature utilisant l'expansion de requête dans un processus sélectif sont de deux sortes : soit elles correspondent à une décision binaire (la requête doit être étendue ou ne doit pas l'être), soit elles décident quel type d'expansion devrait être utilisé. Notre approche suit cette deuxième tendance, en nous focalisant sur les requêtes difficiles. Nous pensons en effet que la nature de la difficulté d'une requête implique des processus d'expansion différents. De plus, comparée à d'autres approches, notre méthode combine des caractéristiques extraites à la fois des requêtes mais également des documents retrouvés et de la collection. Les caractéristiques sont issues de prédicteurs de difficulté des requêtes. Chaque prédicteur se concentre sur un aspect particulier de difficulté de requête ; notre hypothèse est donc que plusieurs mesures de prédiction combinées devraient offrir une caractérisation plus complète et plus fine d'une requête. Les prédicteurs de base sont issus de la littérature et sont soit à base statistique (He et Ounis, 2004 ; Chen *et al.*, 2012), soit à base linguistique (Mothe et Tanguy, 2005). L'efficacité de la méthode que nous proposons est analysée en termes d'efficacité de l'apprentissage, de précision dans les documents obtenus (MAP) et de robustesse.

Le papier est organisé comme suit : la section 2 correspond à une revue de la littérature du domaine. La section 3 présente notre méthode de recherche d'information sélective. Nous détaillons les prédicteurs de difficulté des requêtes que nous utilisons et le modèle de classification de requêtes. La section 4 décrit le cadre d'évaluation. Les résultats et la discussion sont présentés dans la section 5. La section 6 conclut ce papier.

2. Etat de l'art

L'utilisation de différents systèmes ou de différentes configurations de système pour répondre aux besoins des utilisateurs a d'abord été étudié dans le cas des métamoteurs dans lesquels plusieurs résultats sont combinés pour améliorer la liste des documents retrouvés. La combinaison peut correspondre à de la fusion comme dans les fonctions de type "Comb" (Fox et Shaw, 1994) ou servir à favoriser la diversité (Liu *et al.*, 2012 ; Santos *et al.*, 2010). Ces techniques de combinaison de résultats sont généralement appliquées de façon identique sur toutes les requêtes. Par opposition à cela, certaines approches ont étudié l'utilisation d'une fonction dépendante des requêtes.

Ainsi, certaines approches sélectives font l'hypothèse que les requêtes difficiles doivent être traitées différemment des autres. Prédire la difficulté des requêtes devient alors un enjeu important. La prédiction peut être basée sur des indicateurs de pré-traitement ou de post-traitement. De nombreux travaux récents sont consacrés à la définition de prédicteurs. L'*idf moyen des termes de la requête* est un prédicteur simple de pré-recherche : il mesure le pouvoir discriminant des termes de la requête. Le *score de clarté* (clarity score) vise à quantifier le niveau d'ambiguïté d'une requête et correspond à l'entropie relative entre le modèle de langue de la requête et le modèle de langue de la collection de documents (Cronen-Townsend *et al.*, 2002). Le *query scope* (He et Ounis, 2004) mesure le pourcentage de documents de la collection qui contiennent au moins un des termes de la requête. Le *nombre moyen de sens des termes de requête* et la *complexité de la requête* sont d'autres prédicteurs linguistiques de pré-recherche (Mothe et Tanguy, 2005). Le prédicteur post-recherche *gain pondéré d'information* (Zhou et Croft, 2007) mesure l'écart entre la moyenne des scores en haut de la liste des documents retrouvés et l'ensemble du corpus. Yom-Tov et al. proposent une méthode basée sur l'accord entre les résultats de la requête complète et ceux obtenus quand les sous-requêtes sont considérées (Yom-Tov *et al.*, 2005). L'*engagement normalisé de la requête* (Shtok *et al.*, 2009) est un prédicteur post-recherche qui mesure l'écart-type entre les scores des documents retrouvés.

L'expansion sélective a été étudiée comme une application possible de la prédiction de la difficulté des requêtes : le système décide si l'expansion doit être appliquée ou pas, en se basant sur la difficulté de la requête. Dans Amati *et al.* (Amati *et al.*, 2004), la décision est fondée sur le critère *InfoQ*, une fonction issue de la théorie de l'information qui combine la longueur de la requête, l'idf et d'autres caractéristiques. Les auteurs montrent que lorsque l'on considère un modèle d'expansion de requête performant, l'approche sélective peut améliorer la MAP et diminuer le nombre de requêtes sans documents pertinents dans les 10 premiers retrouvés. Cronen-Townsend *et al.* (Cronen-Townsend *et al.*, 2004) comparent le modèle de langue obtenu à partir des documents trouvés en utilisant l'expansion de requête et en utilisant des requêtes non étendues. Le score de la clarté des termes est utilisé pour classer les requêtes en deux groupes : celles qui bénéficieraient de l'expansion et celles qui ne seraient pas étendues.

D'autres approches tentent d'optimiser l'expansion de requête ou d'autres fonctions dans le processus de recherche pour chaque requête prise individuellement. Lv et Zhai (Lv et Zhai, 2009) suggère une méthode d'apprentissage pour prédire le coefficient qui tient compte de la requête initiale et des informations de feedback dans la reformulation PRF. Ils utilisent différentes caractéristiques telles que la discrimination d'une requête (longueur de la requête, l'entropie de la requête et sa clarté), la discrimination des documents de feedback (longueur du feedback, l'entropie, la clarté), et la convergence entre les documents de la requête et du feedback. La régression logistique est utilisée pour apprendre le coefficient d'équilibre. Les auteurs montrent que les trois composantes sont complémentaires et captent différents aspects de l'information. He et Ounis (He et Ounis, 2004) suggèrent une méthode pour choisir parmi plusieurs modèles de pondération des termes en fonction de la requête. Les requêtes

sont représentées par diverses caractéristiques qui sont utilisées pour les regrouper en utilisant une classification ascendante hiérarchique. L'étape d'apprentissage associe le meilleur schéma de pondération des termes à chaque groupe de requêtes. Après apprentissage, une nouvelle requête est d'abord associée à un des groupes existants et le système entrainé est utilisé pour la traiter. Il s'agit d'une méthode pré-recherche puisque les caractéristiques de la requête sont calculées à partir de la requête elle-même et de caractéristiques générales de la collection de documents (longueur de la requête, idf des termes, et clarté/ambiguïté de la requête). Lorsque évaluée sur la collection TREC Robust, la méthode améliore légèrement la MAP lors de l'utilisation des modèles d'expansion si le nombre de requêtes d'entraînement est suffisamment élevé. Chen *et al.* (Chen *et al.*, 2012) considèrent également diverses fonctionnalités pour prédire la performance et appliquer l'expansion de requête sélective. Les auteurs utilisent des caractéristiques telles que la longueur de la requête, la cohérence et l'entropie de la requête, ainsi que des caractéristiques issue de la PRF (probabilités associées aux termes d'expansion) et des caractéristiques de RF explicite (la longueur du document de feedback, l'entropie, le classement, ...). Le modèle prédictif est basé sur une régression logistique qui vise à prédire si la PRF obtiendra de meilleurs résultats que la combinaison de la RF explicite et de la PRF. En utilisant la collection ClueWeb09, la précision du modèle de prédiction est plus de 50%. Le procédé améliore la MAP par rapport à n'importe lequel des procédés où une seule méthode d'expansion est utilisée. Cao *et al.* (Cao *et al.*, 2008) proposent une méthode supervisée pour décider si les termes d'expansion candidats sont bons ou pas, en fonction de leur impact sur la recherche. Le processus est considéré comme un problème de classification des termes qui est résolu en utilisant un SVM .

Dans la méthode que nous proposons, les prédicteurs de difficulté de la requête sont utilisés pour classer les requêtes et associer un modèle d'expansion aux classes de requêtes. Comme dans (He et Ounis, 2004) ou (Chen *et al.*, 2012), nous utilisons les caractéristiques de la requête pour obtenir une décision dépendante des requêtes ; toutefois contrairement à d'autres approches, notre méthode combine des prédicteurs basés sur des aspects linguistiques qui ont par ailleurs été montrés comme corrélés à la difficulté des requêtes (Mothe et Tanguy, 2007). Nous considérons une fonction de sélection qui choisit parmi différentes méthodes d'expansion. Plus précisément, nous considérons la RF et le raffinement de requête. Ces deux méthodes sont orientées vers l'adaptation à l'utilisateur. Cependant, quand une étude avec des utilisateurs n'est pas possible en raison des ressources manquantes et pour des raisons de reproductibilité, il est possible de recourir à une simulation (Lin et Smucker, 2008 ; Zhao et Callan, 2012). Nous simulons donc ces deux types d'expansion. En ce qui concerne la RF nous utilisons la PRF qui considère les premiers x documents retrouvés comme pertinents. Nous simulons le raffinement de requête en considérant la partie descriptive des besoins d'information. Il ne s'agit pas de raffinement au sens interactif, mais plutôt d'un complément d'information sur le besoin de l'utilisateur.

3. Description de la méthode

Une approche sélective en fonction des requêtes traite différemment les requêtes en fonction de critères. Dans notre étude, nous nous sommes centrés sur l'aspect expansion de requête. Plutôt que de savoir s'il faut ou non étendre la requête, nous nous sommes plutôt intéressés à choisir comment étendre au mieux la requête. Notre méthode apprend donc à choisir entre plusieurs alternatives d'expansion de requête. L'apprentissage est réalisé sur un ensemble de requêtes d'entrainement pour lesquelles la meilleure méthode d'expansion est connue. Ces requêtes sont représentées par diverses caractéristiques qui sont en lien avec leur difficulté. L'apprentissage apprend à classer les requêtes en fonction de la meilleure alternative de reformulation. Nous considérons deux alternatives : le raffinement et l'expansion. Notre hypothèse est que les requêtes peuvent être difficiles pour diverses raisons et que donc les différents types de prédicteurs devraient capturer cette diversité. Les méthodes d'expansion pourraient alors refléter une plus ou moins bonne adaptation à la difficulté. Après la phase d'apprentissage, les nouvelles requêtes sont classées sur la base de leurs caractéristiques en suivant le modèle appris, la méthode d'expansion associée est alors utilisée.

3.1. *Raffinement et retour de pertinence*

Notre méthode s'appuie sur deux alternatives d'expansion de requête : le raffinement et le retour de pertinence.

L'expansion de requête par retour de pertinence (RF) utilise des informations issues des documents retrouvés via la requête initiale et jugés comme pertinents par l'utilisateur. Cette méthode implique une interaction ou les jugements de pertinence de la part des utilisateurs. La pseudo RF (PRF) considère les premiers documents retrouvés comme pertinents ; cette méthode est donc complètement automatique. Elle est largement utilisée en recherche d'information. PRF est la première alternative que nous utilisons comme méthode d'expansion.

L'autre alternative que nous utilisons pour étendre une requête est son raffinement. Le raffinement interactif de requête implique des études d'utilisateurs ou l'analyse des traces (log) de connexion sur les requêtes. Des solutions alternatives consistent à simuler le raffinement des requêtes (Zhao et Callan, 2012). Dans notre approche nous considérons un raffinement de la requête par ajout de termes ; ce raffinement est simulé en utilisant la partie descriptive des besoins d'information TREC. Cette partie descriptive simule le raffinement d'une requête en soi, réalisé par un utilisateur, alors on ne considère pas des sous-ensembles de cette partie dans notre étude.

Nous formulons l'hypothèse que les deux types d'expansion (PRF et par raffinement) sont complémentaires et peuvent être utilisés d'une façon sélective, selon les requêtes.

3.2. *Caracteristiques des requêtes*

Les requêtes sont représentées par des caractéristiques qui sont calculées à partir du texte de la requête et de la liste des documents retrouvés par la requête ; ces caractéristiques sont obtenues à partir de 4 prédicteurs de difficulté des requêtes de la littérature.

3.2.1. *Prédicteurs de difficulté des requêtes*

Nous utilisons à la fois des prédicteurs pré-recherche et des prédicteurs post-recherche. Puisque basés sur les termes de la requête seulement, les prédicteurs pré-recherche peuvent être calculés de façon rapide. Toutefois, l'information qu'ils apportent est assez faible car elle est fondée sur peu d'information. Au contraire, les prédicteurs post-recherche s'appuient sur des listes de documents retrouvés et utilisent des informations supplémentaires (scores des documents, rangs). Ce type de prédicteurs implique un processus de recherche afin de traiter la liste de résultats, ils sont donc plus longs à calculer ; en revanche ils ne nécessitent pas de traitement supplémentaire de la collection.

3.2.1.1. Prédicteurs pré-recherche

Le « Nombre de sens de WordNet » (WNS**)** (Mothe et Tanguy, 2005) représente un prédicteur linguistique de pré-recherche, il correspond à une mesure de l'ambiguïté. Il est calculé par le nombre moyen de sens dans WordNet [1] pour tous les termes de la requête q :

$$WNS(q) = \frac{1}{|q|} \sum_{t \in q} senses_t, \qquad [1]$$

où $senses_t$ représente le nombre de synsets de WordNet pour le terme t de la requête q.

La « Fréquence Inverse » (IDF**)** est un prédicteur statistique de pré-recherche et il mesure si un terme est rare ou commun dans le corpus. Sa valeur pour une requête représente la moyenne des IDF pour tous les termes de la requête. L'IDF d'une requête q ($IDF(q)$) est calculé comme suit :

$$IDF(q) = \frac{1}{|q|} \sum_{t \in q} \log_{10} \left(\frac{N}{N_t + 1} \right), \qquad [2]$$

où N est le nombre total de documents dans la collection et N_t le nombre de documents contenant le terme t.

3.2.1.2. Prédicteurs post-recherche

L'écart-type (STD**)** représente un prédicteur post-recherche statistique qui mesure le degré de variation par rapport à la moyenne de la liste des scores attribués

1. `http://wordnet.princeton.edu/`

aux documents retrouvés, correspondant à une requête. C'est une variante de NQC (Shtok *et al.*, 2009), sans normalisation. Pour une requête q et pour les premiers N_q documents retrouvés, le STD est calculé par la formule :

$$STD(q) = \left(\frac{1}{N_q} \sum_{i=1}^{N_q} \Big(score(D_q^i) - \frac{1}{N_q} \sum_{j=1}^{N_q} score(D_q^j) \Big)^2 \right)^{\frac{1}{2}}, \qquad [3]$$

où $score(D_q^i)$ représente le score du $i^{\text{ème}}$ document retrouvé pour q et N_q représente le nombre de documents trouvés par le moteur de recherche pour q . Alors que WNS et IDF sont indépendants de la méthode de recherche, pour STD le modèle de recherche représente un paramètre.

Le feedback de la requête (*QF*) (Zhou et Croft, 2007) est un prédicteur post-recherche qui calcule le chevauchement entre deux listes de documents retrouvés. Les deux listes sont : d'une part celle obtenue par le traitement de la requête initiale (L_1) et d'autre part celle obtenue avec la requête étendue (L_2). A partir de ces deux listes, le chevauchement représente le nombre de documents que ces listes ont en commun. Les listes peuvent être considérées partiellement, à un niveau de coupe x, c'est à dire en ne considérant que les x premiers documents rétrouvés. Afin d'obtenir une mesure entre 0 et 1, le chevauchement est normalisée par division par x. Ainsi QF à plusieurs paramètres : le niveau de coupe et le modèle d'expansion utilisé. Pour une requête q et les deux listes L_1 et L_2 coupées à un niveau x, le QF est calculé comme suit :

$$QF(L_1^x(q), L_2^x(q)) = \frac{|L_1^x(q) \cap L_2^x(q)|}{x}. \qquad [4]$$

Entre les valeurs de l'IDF et l'AP on retrouve une corrélation positive par le coefficient de Pearson (Chifu, 2013). Ainsi un IDF plus grand suggère une requête plus facile. Une corrélation dans le même sens existe pour STD (Chifu, 2013) et QF (Zhou et Croft, 2007). Par contre, la corrélation entre WNS et AP est négative, c'est-à-dire que plus une requête contient des termes polysémiques, plus elle est difficile. QF est normalisé par définition, par contre les autres prédicteurs ne sont pas normalisés. Le mécanisme d'apprentissage est capable de traiter les caractéristiques non normalisées.

3.2.2. *Caractéristiques détaillées de la requête*

A partir de ces quatre facteurs prédictifs de base, nous avons calculé au total 31 caractéristiques. Nous avons choisi des caractéristiques dérivées des prédicteurs de base sous l'hypothèse que plusieurs mesures (min, max, somme, moyenne) peuvent aider le mécanisme d'apprentissage à mieux classifier les requêtes (Guyon et Elisseeff, 2003). Nous ne visons pas l'exhaustivité ni la complétude du modèle à ce stade. Les caractéristiques sont présentées ci-dessous.

L'ambiguïté des termes (6 caractéristiques) : elle est représentée par six caractéristiques dérivées du prédicteur WNS : la moyenne du nombre de sens de la requête (formule 1) est remplacée par le maximum et par la somme. Les caractéristiques sont calculées à la fois pour la requête initiale (T) et la requête raffinée (TD).

La discrimination des termes (8 caractéristiques) : elle est représentée par des éléments issus du prédicteur IDF : la moyenne sur les termes de la requête (formule 2) est remplacée par le minimum, par le maximum et par la somme. Les IDF sont calculés à la fois pour la requête initiale et la requête raffinée.

L'homogénéité des listes de documents (3 caractéristiques) : les caractéristiques de la requête proviennent du prédicteur STD. Nous considérons la valeur des STD pour la liste des documents extraits de la requête initiale et la requête raffinée et leur différence.

La convergence des listes (12 caractéristiques) : les caractéristiques de la requête sont extraites en utilisant le prédicteur QF, en considérant des combinaisons entre les différentes listes pour la requête initiale et la requête étendue ; 4 niveaux de coupe différents sont utilisés (5, 10, 50, 100).

Les combinaisons pré/post-recherche fournissent 2 caractéristiques : le produit entre le WNS de la requête initiale et la différence des STD ainsi que IDF multiplié par la différence des STD.

3.3. *La classification des requêtes*

La classification des requêtes vise à prendre la décision entre la requête complètement étendue (requête raffinée + PRF) et l'expansion PRF seulement. Dans le jeu de données d'entraînement, une requête appartient à la première classe si la précision moyenne (AP) de la requête PRF est supérieure à l'AP de la requête complètement étendue, sinon la requête appartient à l'autre classe. L'apprentissage est basé sur un SVM, qui est une méthode d'apprentissage, bien adaptée à la classification binaire. La première étape utilise une validation croisée de façon à obtenir automatiquement des valeurs optimales des paramètres. Une fois les paramètres fixés, l'algorithme d'apprentissage crée le modèle. Une requête de test est associée à une classe en fonction de sa similitude aux requêtes d'entraînement et le modèle d'expansion correspondant appris est appliqué.

4. Cadre d'évaluation

4.1. *Collections de test et paramètres de la méthode*

Nous avons évalué notre méthode sur trois collections de test de référence, issues de la tâche *ad hoc* de TREC (Text REtrieval Conference)[2] : La collection *Robust* comprend environ 2 Go de documents et 250 besoins d'information, *WT10G* comprend environ 10 Go de documents Web et 100 besoins d'information, et la collection GOV2 comprend environ 25 millions de documents collectés sur les sites .gov et 150

2. `http://mitpress.mit.edu/books/trec`

besoins. TREC propose des besoins d'information (*Topic*) qui sont composés : d'un identifiant, d'un titre, d'un descriptif et d'une partie narrative. La requête qui sera traitée par le système peut être construite par tout moyen à la convenance de l'utilisateur des collections TREC.

Notre méthode de recherche sélective implante deux méthodes d'expansion, mais le moteur de recherche lui-même reste le même. Le moteur de recherche est un point central et les paramètres de configuration de celui-ci peuvent avoir un impact sur les résultats obtenus. Par exemple, certaines caractéristiques des requêtes sont calculées après une première recherche par rapport aux premiers documents retrouvés.

Notre étude s'appuie sur le moteur code source ouvert Indri, basé sur un modèle de langue. Il a été développé par les universités du Massachusetts et Carnegie Mellon[3]. Les listes de documents retrouvés que nous considérons sont composées des 1000 premiers documents.

Les collections ont d'abord été indexées : les mots vides sont supprimés, les mots restants sont racinisés par l'outil Krovetz (Krovetz, 1993). Pour la recherche initiale, nous utilisons le modèle de langue classique, parfois aussi appelé *query likelihood model* (QL) avec un lissage de Dirichlet ($\mu = 1000$). Cette valeur pour μ est cohérente avec plusieurs études de la littérature, comme (Lavrenko, 2009). Pour étendre les requêtes, nous utilisons le modèle de pertinence 3 aussi connu sous le nom de *Relevance Model 3* (RM3) (Lavrenko et Croft, 2001), qui correspond à une interpolation entre le modèle de PRF (RM1) et la requête initiale. Le choix du paramètre pondérant l'importance du modèle de PRF et celui de la requête initiale reste un problème ouvert dans la littérature, nous l'avons fixé à 0.5, donnant ainsi une importance identique à la requête initiale et à sa reformulation.

Quatre variantes de requêtes sont utilisées dans notre approche, en particulier pour calculer les caractéristiques des requêtes :

1) la *requête initiale courte* (dénotée **T**) : le champ titre après suppression des mots vides et racinisation ;

2) la *requête raffinée* de façon manuelle (dénotée **TD**) : elle est simulée en considérant la partie description du besoin d'information, après la suppression des mots vides et racinisation ;

3) la *requête étendue automatiquement* (dénotée **TRF**) : elle est calculée par RM3 (PRF) à partir des résultats obtenus avec la requête initiale courte ;

4) la *requête complètement étendue* qui combine les deux derniers types d'expansion et dénotée **TDRF** : la requête TD est étendue via la PRF.

Le Tableau 1 indique les résultats en termes de MAP obtenus par chacune des variantes. Les résultats sans utiliser PRF sont plus faibles pour toutes les collections. Par rapport à la requête initiale (T), le raffinement (TD) est toujours bénéfique. TDRF

3. `http://www.lemurproject.org/indri/`

donne les meilleurs résultats, sauf pour GOV2, où TRF permet d'obtenir la meilleure MAP.

Tableau 1. *MAP pour chaque collection et pour chaque type de requête*

Collection	MAP T	MAP TD	MAP TRF	MAP TDRF
Robust	0,233	0,260	0,260	0,296
WT10G	0,194	0,215	0,220	0,244
GOV2	0,292	0,294	0,332	0,329

4.2. *Configuration du SVM*

Comme nous l'avons mentionné dans la section 3.2, la matrice des caractéristiques pour le SVM contient les valeurs des caractéristiques calculées à partir de différents prédicteurs de difficulté de requête : il en contient 31 au total.

Concernant le noyau du SVM (Aizerman *et al.*, 1964), nous avons essayé différentes configurations : Gaussien ($k(x, x') = e^{-\gamma||x-x'||^2}$), sigmöide ($k(x, x') = tanh(\gamma x^T x')$) et polynomial ($k(x, x') = (\gamma x^T x')^d$) (jusqu'à 3 degrés). La représentation de la requête dans l'espace des caractéristiques est noté par x. d représente le degré polynomial, et γ est un paramètre d'ajustement. Un autre paramètre d'ajustement est $cost$, qui représente le coût de la violation de la contrainte. Pour les valeurs des paramètres $cost$ et γ, nous avons utilisé une méthode d'échantillonnage automatique sur 10 sous-ensembles (la fonction `tune` dans R[4]). Les meilleures valeurs ont été retenues : $\gamma = 0{,}015625$ et $cost = 4$. Ces valeurs sont inchangées pour les différentes expérimentations faites par la suite.

4.3. *Evaluation*

Pour chacune des collections nous avons utilisé une validation croisée sur 10 sous-ensembles choisis aléatoirement, afin de vérifier la validité de notre modèle d'apprentissage : nous avons utilisé 90% des requêtes pour l'apprentissage et 10% pour le test. Nous avons ensuite calculé la moyenne des résultats en termes de MAP sur l'ensemble des 10 ensembles de test. Suite à l'apprentissage, le SVM est capable d'associer un type d'expansion (TRF ou TDRF) à chaque requête de test.

Par ailleurs, nous nous intéressons à la robustesse du processus. Nous savons qu'en moyenne l'expansion de requête améliore les résultats, mais en revanche, la précision (AP) de certaines requêtes peuvent être dégradées. L'index de Robustesse (IR) (Sakai *et al.*, 2005) est défini comme suit : $IR(q) = \frac{n_+ - n_-}{|q|}$, où q est l'ensemble de requêtes sur lequel est calculé l'IR, n_+ est le nombre de requêtes améliorées, n_- le nombre de

4. `http://www.inside-r.org/packages/cran/e1071/docs/tune`

requêtes dégradées et $|q|$ le nombre total de requêtes. Pour évaluer notre méthode, nous avons calculé l'IR obtenu avec notre proposition et l'IR obtenu avec une approche non sélective, qui étend toutes les requêtes. Cela nous permet de calculer un pourcentage d'amélioration de l'IR.

Les résultats sont présentés dans la section suivante.

5. Résultats et discussion

Comme Guyon et Elisseeff l'indiquent (Guyon et Elisseeff, 2003), avoir des données redondantes peut améliorer le modèle obtenu lors de l'apprentissage. Nous avons testé cette hypothèse, et effectivement, lorsque nous avons utilisé un sous-ensemble des caractéristiques (8 et 22), les résultats ont été dégradés, avec une performance du SVM de 50% lorsque 8 caractéristiques ont été utilisées et de 60% lorsque nous avons utilisé les 22 caractéristiques.

Par ailleurs, nous avons testé plusieurs noyaux pour le SVM, comme présenté dans la section 4.2. Le noyau Gaussien a toujours été le meilleur, et cela de façon assez nette. Par exemple, la précision du SVM dans le cas du noyau polynomial de premier degré a été 33,6% moins bonne que le noyau Gaussien, ainsi l'échec du noyau linéaire suggère qu'il n'y a pas de séparation linéaire entre les deux classes de requêtes. Cette même dégradation (33,6%) a été obtenue pour les noyaux polynomiales de $2^{\text{ème}}$ et $3^{\text{ème}}$ degré. Nous ne présentons dans cet article que les résultats obtenus avec le noyau Gaussien.

Le Tableau 2 indique les améliorations de la MAP, par rapport à une approche non sélective pour les requêtes de test sur les 3 collections. Les valeurs de p-value pour le test statistique (T-test bilatéral apparié) et l'efficacité du modèle sont également indiquées dans ce tableau. Il faut mentionner que dans ce tableau les valeurs de référence correspondent à la moyenne sur les échantillons de test. Elles sont ainsi un peu différentes par rapport aux valeurs référence du Tableau 1.

Tableau 2. *Amélioration en termes de MAP (** : p-value<0,001)*

	Robust	WT10G	GOV2
Référence TRF	0,265	0,234	0,331
Référence TDRF	0,305	0,264	0,334
Résultat avec décision	**0,317**	**0,296**	**0,372**
Amélioration TRF	**19,41%****	**24,65%****	**12,59%****
Amélioration TDRF	**3,90%****	**12,27%****	**11,35%****
Précision du SVM	97,20%	100%	100%
Collins-Thompson : Expansion de référence	0,244	0,183	0,291
Collins-Thompson : Résultats (Robust Feedback)	0,245	0,199	0,300
Collins-Thompson : Amélioration	0,40%	8,73%	3,09%

L'amélioration varie de 3,90% à 24,65% sur la collection de test ; toutes les différences des MAPS sont statistiquement significatifs avec p-value<0,001. Les tests

statistiques ont été conduits entre les populations de référence et les populations des résultats. He et Ounis (He et Ounis, 2007) ont rapporté des amélioration de la MAP jusqu'à $10,83\%$ sur la collection WT10G et jusqu'à 7,40% pour la collection GOV2, en utilisant leur méthode de combinaison de champs pour l'expansion sélective de requête. Par ailleurs, la méthode sélective proposée par Lv et Zhai (Lv et Zhai, 2009) améliore la MAP d'environ 4%. Collins-Thompson (Collins-Thompson, 2009) formule l'expansion de requête comme un problème d'optimisation et la méthode qu'il propose a pour objectif de limiter le compromis risque/récompense de l'expansion. La référence et les résultats qu'il obtient sont présentés dans le Tableau 2. La plus forte amélioration de MAP par rapport à la référence (expansion RF) est de 8,73%, pour la collection WT10G. Notons que la référence qu'il utilisait était plus faible que celle que nous utilisons.

L'expansion sélective est efficace en moyenne, toutefois pour certaines requêtes l'AP peut être dégradée. Par exemple, pour le topic 303 (« hubble telescope achievements ») l'AP est dégradée de 0,1778 à 0,1572 et augmentée de 0,0776 à 0,1949, pour le topic 372 (« native american casino »). Nous avons donc mené une étude de la robustesse. Dans cette étude nous avons calculé l'IR : nous avons comparé la décision d'expansion selon notre méthode à la décision d'expansion unique TDRF. Les résultats sont décrits dans le Tableau 3. Nous indiquons également les résultats obtenus par Collins-Thompson (Collins-Thompson, 2009).

En utilisant les arbres de décision sur le même jeu de données, nous avons établi que les meilleures caractéristiques étaient celles basées sur le prédicteur QF, cela veut dire que le chevauchement entre deux listes représente un facteur discriminant pour la classification, dans notre contexte. Les arbres ont utilisé comme critère l'indice de concentration de Gini et une complexité minimale de 0,0047 (fonction `rpart` dans R[5]).

Tableau 3. *Analyse de la robustesse (** : p-value<0,001)*

	Robust	WT10G	GOV2
IR(TRF)	0,384	0,200	0,413
IR(TDRF)	0,576	0,360	0,187
IR(SVM)	0,752	0,620	0,653
Collins-Thompson : RI	0,377	0,270	0,262
Amélioration IR(TRF)	95,83%**	210%**	58,07%**
Amélioration IR(TDRF)	30,56%**	72,22%**	250%**

Les données de référence pour l'IR sont compatibles avec celles de la littérature (Collins-Thompson et Callan, 2007). L'amélioration de l'IR que nous observons va de 30% jusqu'à 250%. Cela signifie que la méthode sélective est également efficace en termes de robustesse.

5. `http://cran.r-project.org/web/packages/rpart/index.html`

6. Conclusion et perspectives

Nous avons proposé une méthode de RI sélective qui utilise des prédicteurs de difficulté de requêtes pour apprendre à choisir entre plusieurs types d'expansion de requêtes. Notre méthode s'appuie sur une grande variété de caractéristiques sur les requêtes, dont certaines sont linguistiques. La méthode permet d'apprendre efficacement (toujours au dessus de 80%) ; la MAP quant à elle est améliorée de plus de 20% en moyenne avec une bonne robustesse. Les résultats obtenus sont statistiquement pertinents.

La recherche dans le domaine de la prédiction de la difficulté des requêtes permet d'obtenir des résultats de plus en plus performants, même si les prédicteurs définis ne sont pas encore suffisamment utilisés dans des applications concrètes en RI. Nous pensons que le travail sur des approches de RI sélectives sont une piste intéressante dans cette voie.

Concernant les travaux futurs, nous nous intéressons à la définition de nouveaux prédicteurs de difficulté de requêtes ainsi qu'à de nouvelles combinaisons possibles de ces prédicteurs. Par ailleurs, nous avons choisi d'utiliser les SVM comme modèle d'apprentissage, nous pensons tester d'autres techniques comme les arbres de décision qui ont donné de bons résultats dans d'autres domaines. Enfin une dernière piste de travail concerne l'influence relative des différentes caractéristiques dans le modèle.

Remerciements

Ces travaux s'inscrivent dans le cadre du projet CAAS qui a été financé par l'ANR dans l'appel à projet Contint 2010.

7. Bibliographie

Aizerman M. A., Braverman E. M., Rozonoér L. I., « Theoretical foundations of the potential function method in pattern recognition learning », *Automation and Remote Control*, vol. 25, p. 821-837, 1964.

Amati G., Carpineto C., Romano G., « Query Difficulty, Robustness, and Selective Application of Query Expansion », *Advances in Information Retrieval*p. 127-137, 2004.

Arguello J., Diaz F., Callan J., Crespo J.-F., « Sources of evidence for vertical selection », *Proceedings SIGIR'09*, ACM Press, p. 315, July, 2009.

Bigot A., Chrisment C., Dkaki T., Hubert G., Mothe J., « Fusing different information retrieval systems according to query-topics : a study based on correlation in information retrieval systems and TREC topics », *Information Retrieval*, vol. 14, n^{o} 6, p. 617-648, June, 2011.

Buckley C., « Why current IR engines fail », *Proceedings of SIGIR'04*, ACM, p. 584-585, 2004.

Cao G., Nie J.-Y., Gao J., Robertson S., « Selecting good expansion terms for pseudo-relevance feedback », *Proceedings of SIGIR'08*, ACM, p. 243-250, 2008.

Carmel D., Yom-Tov E., « Estimating the Query Difficulty for Information Retrieval », *Synthesis Lectures on Information Concepts, Retrieval, and Services*, vol. 2, n° 1, p. 1-89, January, 2010.

Carpineto C., de Mori R., Romano G., Bigi B., « An information-theoretic approach to automatic query expansion », *ACM Trans. Inf. Syst.*, vol. 19, n° 1, p. 1-27, January, 2001.

Carpineto C., Romano G., « A Survey of Automatic Query Expansion in Information Retrieval », *ACM Computing Surveys*, vol. 44, n° 1, p. 1-50, January, 2012.

Chen C., Chunyan H., Xiaojie Y., « Relevance Feedback Fusion via Query Expansion », *Proceedings of the The 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology - Volume 03*, WI-IAT '12, IEEE Computer Society, Washington, DC, USA, p. 122-126, 2012.

Chifu A., « Prédire la difficulté des requêtes : la combinaison de mesures statistiques et sémantiques (short paper) », *CORIA, Neuchatel, Suisse, 03/04/2013-05/04/2013*, p. 191-200, avril, 2013.

Collins-Thompson K., « Reducing the risk of query expansion via robust constrained optimization », *CIKM*, p. 837-846, 2009.

Collins-Thompson K., Callan J., « Estimation and use of uncertainty in pseudo-relevance feedback », *Proceedings of SIGIR'07*, ACM Press, p. 303, July, 2007.

Cronen-Townsend S., Zhou Y., Croft W. B., « Predicting query performance », *Proceedings of SIGIR'02*, ACM Press, p. 299 - 306, August, 2002.

Cronen-Townsend S., Zhou Y., Croft W. B., « A framework for selective query expansion », *Proceedings of the thirteenth ACM international conference on Information and knowledge management*, CIKM '04, ACM, p. 236-237, 2004.

De Loupy C., Bellot P., « Evaluation of document retrieval systems and query difficultyl », *Proceedings of the Workshop Using Evaluation within HLT Programs : Results and trends*, p. 31-38, 2000.

Fox E. A., Shaw J. A., « Combination of multiple searches », *Proc. TREC-2*, p. 243-249, 1994.

Guyon I., Elisseeff A., « An introduction to variable and feature selection », *The Journal of Machine Learning Research*, vol. 3, p. 1157-1182, March, 2003.

He B., Ounis I., « Inferring query performance using pre-retrieval predictors », *11th International Conference, SPIRE 2004, Proceedings*, Springer Berlin Heidelberg, Padova, p. 43 - 54, 2004.

He B., Ounis I., « Combining fields for query expansion and adaptive query expansion », *Information Processing & Management*, vol. 43, n° 5, p. 1294-1307, September, 2007.

Krovetz R., « Viewing Morphology As an Inference Process », *Proceedings of SIGIR'93*, ACM, p. 191-202, 1993.

Lavrenko V., *A Generative Theory of Relevance*, vol. 26 of *The Information Retrieval Series*, Springer, Berlin, 2009.

Lavrenko V., Croft W. B., « Relevance based language models », *Proceedings of SIGIR'01*, ACM Press, p. 120-127, September, 2001.

Lin J. J., Smucker M. D., « How do users find things with PubMed ? : towards automatic utility evaluation with user simulations », *SIGIR*, ACM, p. 19-26, 2008.

Liu H., Wu Z., Hsu D. F., « Combination of Multiple Retrieval Systems Using Rank-Score Function and Cognitive Diversity », *in* L. Barolli, T. Enokido, F. Xhafa, M. Takizawa (eds), *AINA*, IEEE, p. 167-174, 2012.

Lv Y., Zhai C., « Adaptive relevance feedback in information retrieval », *Proceedings of the 18th ACM conference on Information and knowledge management*, CIKM '09, ACM, p. 255-264, 2009.

Mothe J., Tanguy L., « Linguistic features to predict query difficulty », *ACM Conference on research and Development in Information Retrieval, SIGIR, Predicting query difficulty - methods and applications workshop , Salvador de Bahia, Brésil, 15/08/05-19/08/05*, ACM, p. 7-10, 2005.

Mothe J., Tanguy L., « Linguistic Analysis of Users' Queries : Towards an Adaptive Information Retrieval System », *2007 Third International IEEE Conference on Signal-Image Technologies and Internet-Based System*, IEEE, p. 77-84, December, 2007.

Peng J., Macdonald C., Ounis I., « Learning to select a ranking function », *Proceedings of the 32nd European conference on Advances in Information Retrieval*, ECIR'2010, Springer-Verlag, Berlin, Heidelberg, p. 114-126, 2010.

Sakai T., Manabe T., Koyama M., « Flexible pseudo-relevance feedback via selective sampling », *ACM Transactions on Asian Language Information Processing*, vol. 4, n° 2, p. 111-135, June, 2005.

Santos R. L. T., Macdonald C., Ounis I., « Selectively diversifying web search results », *in* J. Huang, N. Koudas, G. J. F. Jones, X. Wu, K. Collins-Thompson, A. An (eds), *CIKM*, ACM, p. 1179-1188, 2010.

Shtok A., Kurland O., Carmel D., « Predicting Query Performance by Query-Drift Estimation », *Advances in Information Retrieval Theory, Proceedings of ICTIR 2009*, vol. 5766 of *Lecture Notes in Computer Science*, Springer, p. 305-312, september, 2009.

Yom-Tov E., Fine S., Carmel D., Darlow A., « Learning to estimate query difficulty », *Proceedings of SIGIR'05*, ACM Press, p. 512, August, 2005.

Zhao L., Callan J., « Automatic term mismatch diagnosis for selective query expansion », *Proceedings of SIGIR'12*, ACM Press, p. 515, August, 2012.

Zhou Y., Croft W. B., « Query performance prediction in web search environments », *in* W. Kraaij, A. P. de Vries, C. L. A. Clarke, N. Fuhr, N. Kando (eds), *Proceedings of SIGIR'07*, ACM, p. 543-550, 2007.