Expansion de requêtes par apprentissage [1]

Ahlem Bouziri* — **Chiraz Latiri** * — **Eric Gaussier****

* *Université de la Manouba (ISAMM)- LIPAH (FST), Tunisie*
** *Université Joseph Fourrier-Laboratoire d'Informatique de Grenoble, France*

RÉSUMÉ. *Cet article propose une approche d'expansion automatique de requêtes par apprentissage. L'expansion de requêtes se fait par l'ajout de termes provenant de règles d'association entre termes. Le problème d'expansion de requêtes est modélisé comme un problème de classification superviséE qui vise à déterminer les règles d'association les plus adaptées pour enrichir une requête donnée. Un ensemble de données d'entraînement est construit en utilisant un algorithme d'exploration de règles d'association pertinentes, basé sur les algorithmes génétiques. La classification se fait par la méthode de l'arbre de décision et par la méthode Random Forest. Les expérimentations sont menées sur la collection de textes en français SDA95 de la campagne d'évaluation CLEF 2003. Les résultats montrent une amélioration des performances de la tâche RI*

ABSTRACT. *We propose in this paper a learning query expansion approach using association rules. The query expansion problem is modeled as a supervised classification problem which aims at identifying the appropriate set of association rules to expand a given query. A training data set is generated using a GA based exploring algorithm of the association rules space. Classification is made by the method of the decision tree and the Random Forest method . The experiments are conducted on the French texts SDA95 collection of CLEF evaluation campaign 2003. The results show an improvement in task performance IR.*

MOTS-CLÉS : *règles d'association, expansion de requêtes, apprentissage supervisé.*

KEYWORDS: *Association rules, query expansion, supervised learning.*

1. Traduction étendue d'un article publié en anglais à KDIR 2015 (Bouziri *et al.*, 2015)

1. Introduction

La Recherche d'Information (RI) étudie le processus d'adéquation entre la requête d'un utilisateur et une collection de documents, dont le résultat est souvent un sous-ensemble de documents pertinents contenant les mêmes termes de la requête originelle. Le modèle classique de RI (Salton et McGill, 1983) consiste à attribuer, à chaque document d'une collection, des termes d'indexation, dits index du document, limitant les requêtes à l'ensemble global des termes de l'index, et utilisant des mesures de correspondance entre les requêtes et les documents. Une des difficultés rencontrées au cours d'une session de recherche est liée aux choix des termes d'interrogation. En effet, dans bien des cas, les termes de la requête, exprimée par l'utilisateur, ne correspondent pas exactement aux descripteurs des documents retenus par le modèle d'indexation. De ce fait, afin d'avoir des documents pertinents, l'utilisateur est contraint d'utiliser le "vocabulaire de description du document" propre au système. Face à cette contrainte, il est possible de faire appel à la technique d'expansion de requêtes (Buckley *et al.*, 1994) afin d'améliorer la correspondance requête/document, et ce en étendant la requête par des termes additionnels, corrélés à ceux de la requête originelle. Intuitivement, l'apport d'une telle technique ne se réduit pas à l'amélioration du rappel en récupérant des documents pertinents qui ne peuvent pas être trouvés par la requête utilisateur, mais également à améliorer la précision des documents trouvés en les plaçant en haut de la liste des documents pertinents.

Nous proposons dans cet article une approche d'expansion de requêtes par apprentissage qui se base sur les règles d'association entre termes. Un processus exploratoire basé sur les algorithmes génétiques explore l'espace des règles d'association entre termes à la recherche des meilleurs termes d'expansion et génère en parallèle des instances d'entrainement qui serviront pour construire un classifieur. Intuitivement, une règle d'association traduit la probabilité d'avoir les termes de la conclusion dans un document, sachant que ceux de la prémisse y sont. Ainsi, l'utilisation de telles dépendances dans un processus d'expansion de requêtes améliore sensiblement la pertinence d'un SRI, car elles reflètent des corrélations fortes et implicites découvertes à partir de la collection de documents. Toutefois, face au nombre très important de règles d'association entre termes qui peuvent être découvertes à partir d'une collection de documents, nous proposons un nouveau processus d'expansion automatique de requêtes moyennant la base générique minimale $\mathcal{MGB}$ déployée dans (Latiri *et al.*, 2012).

2. Expansion de requêtes : revue de la littérature

La problématique d'expansion de requêtes a été largement abordée par la communauté RI depuis deux décennies (Buckley *et al.*, 1994 ; Ruthven et Lalmas, 2003 ; Kumaran et Allan, 2008 ; Carpineto et Romano, 2012).

Certaines approches d'expansion de requêtes se basent sur des techniques de fouille de texte et utilisent des connaissances dérivées à partir des collections de textes. Elles observent généralement la régularité des termes dans un contexte déterminé

d'une collection de textes. Elles sont basées sur l'hypothèse qui stipule que l'emploi de deux termes en co-occurrence est l'expression d'une relation sémantique entre eux (Rijsbergen, 1979). L'avantage de ces approches est qu'elles sont faciles à mettre en oeuvre tout en étant indépendantes du corpus. Parmi les premiers travaux relatifs à cette classe d'approches, Grefensette (Grefenstette, 1992) contribue par une approche syntaxique d'extraction de contextes de mots à partir des corpus textuels pour produire la liste des mots reliés à n'importe quel mot du corpus. D'autres approches s'appuient sur une analyse statistique des collections par l'extraction de règles d'association entre termes, afin d'ajouter des termes voisins à la requête originelle (Tangpong et Rungsawang, 2000 ; Haddad *et al.*, 2000). Les associations sont généralement basées sur la co-occurrence des termes dans les documents (Rungsawang *et al.*, 1999). L'usage d'une telle technique a montré que les liens inter-termes renforcent la notion de pertinence des documents par rapport aux requêtes (Latiri *et al.*, 2012).

D'autres travaux se sont intéressés à l'optimisation de l'expansion en utilisant des métaheuristiques tels que les algorithmes génétiques. Dans (Boughanem et Tamine, 2002), les auteurs présentent une approche d'expansion fondée sur une utilisation combinée de la stratégie d'injection de pertinence et des techniques avancées de l'algorithme génétique. La population est représentée par un ensemble de niches, renouvelée à chaque génération. La fonction d'adaptation est basée à la fois sur le jugement de pertinence des utilisateurs et sur un modèle de fonction statistiquement corrélé aux mesures de taux de rappel/precision. Les opérateurs génétiques sont appliqués de manière restreinte aux niches et non pas de manière uniforme sur toute la population. La sélection adoptée dans cette approche est fondée sur l'espérance mathématique qui permet de générer pour chaque individu de la population un nombre de clones dépendant de sa valeur d'adaptation. Le croisement est basé sur le poids des termes. La mutation consiste essentiellement à exploite les termes présents dans les documents pertinents afin d'ajuster les valeurs des gènes correspondants dans les requêtes sélectionnées pour la mutation. Par ailleurs, les auteurs de (Araujo *et al.*, 2010) proposent une méthode basée sur les algorithmes génétiques pour résoudre les problèmes relatifs à la sélection de termes adéquats pour la reformulation de requêtes utilisant la structure morphologique des requêtes.

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. Chifu et Mothe (Chifu et Mothe, 2014) présentent une méthode d'expansion sélective qui se base sur des prédicteurs de difficulté des requêtes. Le modèle de décision, appris par un SVM, permet de prédire la nature de l'expansion selon la difficulté de la requête.

Dans un couplage entre RI et fouille de textes, une approche d'expansion automatique de requêtes basée sur les règles d'association a été proposée dans (Latiri *et al.*, 2012). Cette approche commence par dériver, dans un premier temps, la base générique $\mathcal{MGB}$ de règles d'association non-redondantes entre termes à partir d'une collection de documents, et de l'utiliser, dans un deuxième temps, pour étendre la re-

quête originelle de l'utilisateur. Chaque requête originelle de la collection est ainsi étendue en injectant tous les termes qui apparaissent dans les conclusions des règles d'association, qui ont dans leurs prémisses respectives les termes de la requête originelle. Les expérimentations menées dans le cadre de l'évaluation de cette approche montrent, pour les différentes collections testées, une amélioration de la pertinence système en terme de MAP, réalisée avec les requêtes étendues par les règles d'association de la base générique $\mathcal{MGB}$ (Latiri *et al.*, 2012).

Dans le présent article, nous nous intéressons au raffinement de cette approche par une meilleure sélection des règles d'association candidates à l'expansion. En effet, nous partons de l'hypothèse que si une expansion de requêtes par injection systématique des termes des conclusion de toutes les règles d'association qui ont dans leurs prémisses respectives au moins un terme de la requête originelle, s'est avérée efficace ; alors elle serait encore plus efficiente si seules les règles d'association qui améliorent la pertinence de la recherche sont considérées.

3. Définitions et formalisme de base

3.1. *Les algorithmes génétiques*

Les algorithmes génétiques, initiés dans les années 1970 par John Holland, sont des algorithmes d'optimisation s'appuyant sur des techniques dérivées de la génétique et des mécanismes d'évolution de la nature : croisement, mutation, sélection. Un algorithme génétique recherche le ou les extrema d'une fonction définie sur un espace de données. Pour l'utiliser, on doit disposer des cinq éléments suivants :

– *Un principe de codage de l'élément de population.* Cette étape associe à chacun des points de l'espace d'état une structure de données. Elle se place généralement après une phase de modélisation mathématique du problème traité. La qualité du codage des données conditionne le succès des algorithmes génétiques. Les codages binaires ont été très utilisés à l'origine. Les codages réels sont désormais largement utilisés, notamment dans les domaines applicatifs pour l'optimisation de problèmes à variables réelles.

– *Un mécanisme de génération de la population initiale.* Ce mécanisme doit être capable de produire une population d'individus non homogène qui servira de base pour les générations futures. Le choix de la population initiale est important car il peut rendre plus ou moins rapide la convergence vers l'optimum global.

– *Une fonction à optimiser.* Celle-ci retourne une valeur appelée *fitness* ou fonction d'évaluation de l'individu.

– *Des opérateurs permettant de diversifier la population au cours des générations et d'explorer l'espace d'état.* L'opérateur de croisement recompose les gènes d'individus existant dans la population, l'opérateur de mutation a pour but de garantir l'exploration de l'espace d'états.

– *Des paramètres de dimensionnement* : taille de la population, nombre total de générations ou critère d'arrêt, probabilités d'application des opérateurs de croisement et de mutation.

3.2. *Les règles d'association entre termes*

Dans ce travail, nous nous basons sur une base générique de règles d'association entre termes appelée $\mathcal{MGB}$ proposée dans (Latiri *et al.*, 2012). La spécificité de cette base est qu'elle est compacte, dans le sens où elle englobe un noyau minimal de règles d'association approximatives et exactes non-redondantes entre termes. La caractéristique clé des règles dérivées est qu'elles ont des prémisses minimales illustrées par les générateurs minimaux et des conclusions maximales. De ce fait, les règles d'association s'avèrent intéressantes dans le contexte d'expansion de requêtes étant donné que ces prémisses maximales offrent plus de termes candidats à l'expansion. Leur déploiement en expansion de requêtes en RI a été également détaillé dans (Latiri *et al.*, 2012).

Une règle d'association entre termes R est une implication de la forme $R : T_1 \Rightarrow T_2$, où T_1 et T_2 sont deux sous-ensembles distincts de l'ensemble de tous les termes distincts des documents de la collection , et $T_1 \cap T_2 = \emptyset$. T_1 et T_2 sont, respectivement, appelés la *prémisse* et la *conclusion* de R. La règle R est ainsi dérivée à partir de $T_1 \cup T_2$ (Latiri *et al.*, 2012).

Le *support* de la règle $R : T_1 \Rightarrow T_2$ représente le nombre de documents d de la collection C qui contiennent tous les termes t de $T_1 \cup T_2$. Il est défini comme suit :

$$\textit{Supp}(R) = \textit{Supp}(T_1 \cup T_2) = |\{d | d \in \mathcal{C} \ \wedge \ \forall\, t \in T_1 \cup T_2 : t \in d\}| \qquad [1]$$

Alors que sa *confiance* est calculée comme suit :

$$\textit{Conf}(R) = \frac{\textit{Supp}(T)}{\textit{Supp}(T_1)} = \frac{\textit{Supp}(T_1 \cup T_2)}{\textit{Supp}(T_1)} \qquad [2]$$

Une règle d'association R est dite *valide* si sa valeur de confiance, *i.e.*, *Conf*(R), est supérieure ou égale à un seuil prédéfini noté *minconf*. Ce seuil de confiance minimal est utilisé pour exclure les règles dites *non valides*. Par ailleurs, le seuil de support minimal *minsupp* est utilisé pour écarter les règles d'association dérivées à partir du termset T et qui ne sont pas suffisamment fréquentes, *i.e.*, les règles ayant un support *Supp*(T) $<$ *minsupp*.

Dans la littérature, deux types de règles d'association sont définis à savoir : *les règles exactes* (avec une confiance égale à 1) et *les règles approximatives* (avec une confiance strictement inférieure à 1) (Zaki, 2004).

4. Apprentissage de l'expansion de requêtes par règles d'association

4.1. *Aperçu général de l'approche*

Notre système est formé de deux composants principaux comme le montre la Figure 1. Un premier composant est dédié à la construction des exemples d'entraînement. Le deuxième composant est chargé de la construction d'un modèle de prédiction à partir des exemples d'apprentissages et en utilisant un algorithme de classification supervisée. Etant donné une requête q et l'ensemble des règles d'association qui lui correspondent RA_q, le modèle de prédiction identifie RA_q^+ le sous ensemble de règles d'association qui sont à utiliser pour générer la requête étendue Eq.

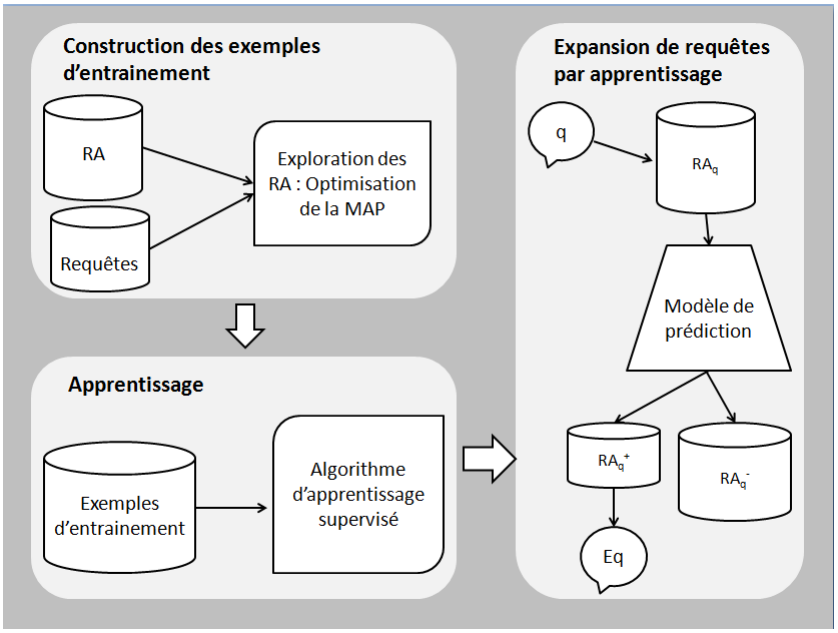

Figure 1. *Schéma général de l'approche*

4.2. *Définition et modélisation du problème*

Nous définissons le problème d'expansion de requêtes comme suit : étant donné une requête originelle $q := \{t_1, t_2, \ldots, t_n\}$, et l'ensemble des règles d'association qui lui sont relatives AR_q ; il s'agit de trouver les règles d'association de AR_q^+ dont les termes des conclusions sont les plus adaptés à enrichir q et restituer des documents qui répondent aux besoins de l'utilisateur.

Nous modélisons ce problème comme un problème de classification supervisée dans lequel il s'agit de prédire si une règle devrait être utilisée pour enrichir une re-

quête donnée. La classification consiste à définir une fonction $f : \mathcal{X} \rightarrow \mathcal{Y}$ qui associe pour chaque observation x sa classe y.

Dans le contexte de l'expansion de requêtes utilisant les règles d'association :

– $\mathcal{X} \subseteq \mathbb{R}^d$ constitue l'ensemble des vecteurs à d dimension représentant l'espace des observations

– $x_{ij} \in \mathcal{X}$ est le vecteur des attributs d'une requête q_i et de la règle d'association AR_j

– $\mathcal{Y} = \{0, 1\}$ est ensemble des classes possibles

– $y_{ij} \in \mathcal{Y}$ est la classe associée à x_{ij}.

- $y_{ij} = 1$ si la règle d'association AR_j est selectionnée par l'algorithme d'optimisation pour enrichir la requête q_i

- $y_{ij} = 0$ sinon.

Le vecteur d'attributs représenté par des attributs calculés à partir du texte de la requête et des termes de la conclusion de la règle d'association. Ces attributs sont au nombre de 10 et sont calculés à partir de mesures statistiques.

4.2.1. *Attributs basés sur la fréquence de documents*

La fréquence de documents (DF) d'un terme est une mesure statistique qui renseigne sur le nombre de documents qui contiennent le terme en question. Sa valeur pour une requête représente la moyenne du DF de tous les termes de la requête. La valeur $DF(q)$ d'une requête q est calculée selon l'équation 3.

$$DF(q) = \frac{1}{|q|} \sum_{i=1}^{|q|} \frac{|\{d_j, t_i \in d_j\}|}{|C|} \quad [3]$$

Nous utilisons également

$$iDF(q) = \frac{1}{|q|} \sum_{i=1}^{|q|} \log \frac{|C|}{|\{d_j, t_i \in d_j\}|} \quad [4]$$

Pour une règle d'association, les DF et iDF sont calculés selon le même principe, et représentent la moyenne des DF (resp. iDF) des termes de la conclusion de la règle.

4.2.2. *Attributs basés sur la fréquence des termes*

Le deuxième type d'attributs que nous prenons en compte dans les instances d'entraînement concerne la fréquence d'un terme dans les documents de la collection. La fréquence ($TF(t, d)$) correspond au nombre d'occurrences du terme t dans le document d. Pour chaque terme de la requête ou de la conclusion de la règle d'association, nous utilisons la moyenne des fréquences dans tous les documents et nous la notons

$ATF(t_i)$ pour le terme t_i. Pour une requête, la fréquence des termes est la moyenne des ATF de tous les termes de la requête et est calculée selon la formule 5.

$$TF(q) = \frac{1}{|q|}\sum_{i=1}^{|q|} ATF(t_i) \quad [5]$$

$$ATF(t_i) = \frac{1}{|C|}\sum_{j=1}^{|D|} TF(t, d_j)$$

Dans le calcul de la mesure $TF(q)$, nous nous basons sur une moyenne des fréquences de chaque terme dans tous les documents $ATF(t)$. Afin de tenir compte de la dispersion des différentes fréquences d'un terme dans chaque document par rapport à la moyenne ATF, nous introduisons la mesure VTF(q) qui représente la moyenne des variances des fréquences des termes de la requête. Le calcul de $VTF(q)$ se fait comme l'indique la formule 6.

$$VTF(q) = \frac{1}{|q|}\sum_{i=1}^{|q|} \frac{1}{|C|-1}\sum_{j=1}^{|C|}(TF(t_i, d_j) - ATF(t_i))^2 \quad [6]$$

4.2.3. *Attributs basés sur les caractéristiques des règles d'association*

Nous utilisons 6 attributs dans le but de caractériser les règles d'association, à savoir :

– Le nombre de termes dans la conclusion de la règle,

– Le nombre de termes dans la prémisse de la règle,

– La confiance de la règle donnée dans l'équation (2),

– Le support de la règle donné dans l'équation (1).

– La proportion des termes de la requêtes qui se retrouvent dans la prémisse.

– Une mesure de l'importance des termes de la conclusion dans la base des règles d'association

4.3. *Génération des exemples d'entraînement*

Les exemples d'entraînement de la base d'apprentissage correspondent dans notre cas à des couples (requêtes , règles d'association). Pour chaque requête, les instances générées sont au nombre de règles d'association relatives à cette requête. Une instance est de classe positive si la règle d'association a été retenue pour l'expansion de la requête, elle est de classe négative sinon. Les règles d'association retenues pour l'expansion sont choisies lors d'un processus exploratoire qui se base sur un algorithme

génétique. Il prend en entrée une requête originelle et l'ensemble des règles d'association candidates à l'expansion pour cette requête. Il retourne une requête étendue optimisée et génère les instances d'apprentissage relatives à l'expansion optimale. Les principes de ce processus exploratoire sont décrit dans ce qui suit.

4.3.1. *Individu et Population*

Un individu ou chromosome représente une requête étendue candidate.

Les gènes de l'individu sont les termes candidats à l'expansion issus à partir des règles d'association entre termes de la base générique $\mathcal{MGB}$. Pour ce faire, chaque terme de la requête originelle est traité individuellement en le cherchant dans les prémisses minimales des règles valides. Les termes des conclusions maximales de toutes les règles ayant comme prémisse le ou les termes de la requête originelle sont retenus comme candidats à l'expansion. La longueur d'un chromosome correspond au nombre de termes candidats à l'expansion.

Étant donnée une requête originelle Q = $\{t_1,\ldots,t_n\}$, l'ensemble des termes candidats à l'expansion obtenu à partir des règles d'association de la base $\mathcal{MGB}$, noté TC, est exprimé comme suit (Latiri *et al.*, 2012) :

$$\begin{aligned} \forall R : T_1 \Rightarrow T_2, \text{une règle non-redondante} \in \mathcal{MGB}; \\ si T_1 \subseteq \text{Q} \\ alors \text{TC} = \text{TC} \cup T_2. \end{aligned} \qquad [7]$$

L'équation (7) signifie que si la prémisse de R est contenue dans Q, les termes de la conclusion sont candidats à l'expansion.

Nous adoptons un codage binaire qui renseigne sur l'apparition ou non d'un terme dans la requête étendue candidate.

4.3.2. *Population initiale*

Chaque individu de la population initiale correspond à une requête étendue formée par les termes de la requête originelle auxquels sont ajoutés les termes des conclusions d'une règle d'association de la base $\mathcal{MGB}$ dont les termes de la prémisse sont des termes de la requête originelle. Pour construire la population initiale, on commence par filtrer la base générique d$\mathcal{MGB}$ pour ne garder que les règles d'association dont les termes des prémisses sont des termes de la requête originelle.

Le nombre de ces règles détermine la taille de la population initiale. Le nombre de gènes des individus correspond au nombre de termes différents trouvés dans les conclusions de ces règles. Ces termes sont candidats à l'expansion.

4.3.3. *Fonction d'évaluation*

L'objectif derrière l'expansion de requête est d'améliorer la pertinence de la tâche RI. La force d'un individu représente la capacité de la requête étendue correspondante à produire des documents pertinents. Nous utilisons la mesure d'évaluation des SRI, la MAP comme fonction d'évaluation. Pour ce faire, chaque individu/requête est soumis à un SRI expérimental (Terrier) pour en évaluer la force de l'individu.

4.3.4. *Opérateurs génétiques*

4.3.4.1. Sélection

Cet opérateur permet aux individus d'une population de survivre, de se reproduire ou de mourir. En règle générale, la probabilité de survie d'un individu sera directement reliée à son efficacité relative au sein de la population. Dans cette implémentation, nous optons pour une sélection élitiste qui consiste à sélectionner les n individus parents dont on a besoin pour la nouvelle génération en prenant les n meilleurs individus de la population courante après l'avoir triée de manière décroissante selon la valeur la fonction d'évaluation de ses individus.

4.3.4.2. Croisement

Le rôle fondamental de l'opérateur de croisement est de permettre la recombinaison des informations présentes dans le patrimoine génétique de la population. L'opérateur de croisement favorise l'exploration de l'espace de recherche. Le croisement permet dans notre cas de produire de nouvelles requêtes étendues et enrichies avec de nouveaux termes à partir des requêtes/individus de la population courante. Nous optons pour un opérateur de croisement qui produit un seul enfant à partir de deux parents. La requête/ individu enfant hérite les termes d'expansion de ses deux parents.

4.3.4.3. Remplacement

A chaque itération de l'AG, les nouveaux individus remplacent leur parents.

5. Évaluation expérimentale de l'approche d'expansion

5.1. *Cadre d'évaluation*

L'évaluation expérimentale a été menée avec le SRI TERRIER, en utilisant la collection de textes français SDA-95 de la campagne CLEF 2003. Quelques caractéristiques statistiques de la collection français *French SDA 95*, notée dans la suite de l'article *SDA-95* sont données dans le Tableau 1.

Tous les champs de la requête, *i.e.,* titre, champs descriptifs et narratifs, sont utilisés lors du processus d'expansion. Nous avons testé notre approche d'expansion avec le schémas de pondération OKAPI BM25 qui est une méthode d'ordonnancement de la méthode OKAPI, la plus connue des méthodes probabilistes, et ayant pour but de

Collection	SDA95
Nombre de documents	42615
Nombre de phrases	523320
Nombre de termes	6664000
Nombre de requêtes	60

Tableau 1. *Caractéristiques de la collection* FRENCH SDA 95 (SDA-95) *de CLEF 2003.*

construire un modèle probabiliste qui prend en compte la fréquence des termes ainsi que la taille des documents (Jones *et al.*, 2000).

5.2. *Pré-traitements et bases comparatives*

Nous avons procédé à un ensemble d'évaluations afin de déterminer des bases de comparaison pour évaluer la performance du processus d'expansion automatique de requêtes par apprentissage proposé dans cet article.

5.2.1. *Évaluation des résultats pour les requêtes sans expansion*

Il s'agit de déterminer la base d'évaluation comparative (baseline). La pertinence des documents retournés est estimée selon les mesures d'évaluation suivantes :

– Précision de la requête originelle (Q) à 11 points de rappel (P@11).

– Les précisions à P@5, P@10, P@15, et P@30 documents pertinents restitués.

– La précision moyenne MAP (Mean Average Precision). Comme la courbe de la précision à 11 points de rappel, la MAP définit la performance globale d'un SRI.

5.2.2. *Évaluation des résultats de recherche pour les requêtes étendues avec PRF*

La pseudo injection de pertinence (ou PRF de l'anglais Pseudo Relevance Feed Back) est une méthode d'expansion complètement automatique largement utilisée en recherche d'information. Elle considère les premiers documents retrouvés via la requête initiale comme pertinents et utilise les informations issues de ces documents pour l'expansion. Nous avons utilisé la méthode PRF implémentée dans Terrier avec les paramètres par défaut (T=10, D=3) pour établir une comparaison avec une méthode d'expansion basique en RI.

5.2.3. *Génération de la base générique des règles d'association $\mathcal{MGB}$*

Lors de la génération de la base générique $\mathcal{MGB}$ (Latiri *et al.*, 2012), nous avons fait varier les seuils minimaux de support et de confiance, *i.e.,* $minsupp$ et $minconf$. Rappelons que ces seuils sont définis pour éliminer, respectivement, les règles très rares et celles qui ne sont pas valides. Nous avons choisi de générer une base géné-

rique $\mathcal{MGB}$ avec un large nombre de règles d'association entre termes afin de diversifier les termes candidats à l'expansion et de garantir plus d'efficacité au processus d'expansion. Pour ce faire nous avons opté, moyennant une étude de la distribution de Zipf des termes du vocabulaire, pour une valeur de $minsupp$=0.003 et pour une valeur de $minconf$=0.1.

5.2.4. *Expansion de requêtes en utilisant $\mathcal{MGB}$*

Afin de montrer l'intérêt de notre approche, nous avons procédé à une expansion "aveugle" des requêtes en utilisant la base générique des règles d'association $\mathcal{MGB}$. Dans cette expansion, nous ajoutons les termes des conclusions de toutes les règles de $\mathcal{MGB}$ ayant au moins un terme de la prémisse dans la requête.

5.2.5. *Génération d'exemples d'entraînement*

Le processus d'optimisation de l'expansion de requêtes basé sur les algorithmes génétique (*c.f*, section 4.3) est utilisé pour générer à la fois les requêtes étendues optimisées et les instances d'entraînement correspondantes aux 60 requêtes de la collection.

Il est important de noter que le nombre total d'instances générées pour les 60 requêtes est de 14495 avec 14058 instances négatives contre seulement 437 instances positives.

5.3. *Protocole expérimental*

Nous avons procédé à une validation croisée à 4 échantillons de test. Les instances d'entraînement correspondants aux 60 requêtes de la collection sont divisées en 4 sous-ensembles, chacun comportant les instances relatives à 15 requêtes. Le processus d'expérimentation se fait en quatre itérations, à chaque fois 3 sous-ensembles d'instances sont considérées pour construire le classifieur et le quatrième est utilisé pour la validation du modèle. Nous faisons de sorte que les 4 sous-ensembles soient testés. Les résultats de classification obtenus dans les quatre phases de test sont fusionnés en un ensemble de 60 requêtes étendues par apprentissage qui est soumis à une évaluation de la recherche.

Le pourcentage de classification correcte (PCC) est une mesure de performance largement utilisée dans la littérature pour l'évaluation des classifieurs. Dans certains cas, cette mesure s'avère peu ou pas du tout significative. Cette inefficacité s'exprime, surtout, dans le cas où il existe un déséquilibre important dans la distribution des classes comme c'est le cas dans nos instances d'entraînement. En effet, un classifieur qui affecte systématiquement les instances à la classe la plus représentée dans la base, aura un PCC correspondant à la proportion de cette classe qui dépasse les 90%. Par conséquent, ce classifieur pourra être considéré comme étant un excellent classifieur si la majorité des instances appartiennent à la classe en question. Ceci est faux, car, réellement, ce classifieur ne dispose d'aucun pouvoir prédictif. Nous avons eu recours

Tableau 2. *Performance des classifieurs J48*

Itération	PCC	PVP	PVN
1	96,00%	11,5%	98,4%
2	97,08%	7,4,52%	99,7%
3	97,09%	7,7%	99,7%
4	95,32%	8,0%	98,9%
Moyenne	96,37%	8,65%	99,17%

Tableau 3. *Performance des classifieurs Random Forest*

Itération	PCC	PVP	PVN
1	97,03%	7,07%	99,55%
2	97,08%	4,62,52%	99,75%
3	97,17%	1,92%	99,94%
4	95,88%	5,35%	99,56%
Moyenne	96,79%	4,74%	99,7%

à des mesures plus précises qui représentent les proportions des instances correctement prédites parmi les instances positives (PVP) et négatives(PVN). Un classifieur sera d'autant plus performant qu'il sera capable de concilier un PVP et un PVN élevés.

5.3.1. *Classifieurs obtenus par la méthode de l'arbre de décision*

A chaque itération de la validation croisée, la méthode J48 de Weka qui implémente la méthode de l'arbre de décision C4.5 (Quinlan, 1987) est utilisée pour générer un classifieur. Les valeurs des paramètres retenus sont ceux par défaut. Les performances des quatre classifieurs obtenus sont résumées dans la Table 2

5.3.2. *Classifieurs obtenus par la méthode Random Forest*

De même que pour la méthode de l'arbre de décision, à chaque itération de la validation croisée un classifieur est construit par la méthode $RandomForest$ de Weka. Les performances des quatre classifieurs obtenus sont résumées dans la Table 3

5.4. *Résultats expérimentaux et discussion*

Nous avons utilisé les mesures de précision P@5, P@10, P@20 et P@30 documents qui sont respectivement, la précision moyenne aux 5, 10, 20 et 30 premiers documents retournés et la MAP (Mean Average Précision), sur l'ensemble des 60 requêtes. Pour chaque requête, les 1 000 premiers documents sont renvoyés par le SRI expérimental et les précisions moyennes sont calculées pour mesurer la pertinence

Tableau 4. *Résultats expérimentaux pour la collection SDA-95*

	MAP	P@5	P@10	P@20	P@30
Baseline	0,355	0,380	0,280	0,216	0,176
BE	0,282	0,286	0,221	0,153	0,128
Δ BE	-20,5%	-24,7%	-21%	-29,1%	-27,2%
PRF	0,358	0,392	0,296	0,212	0,180
Δ PRF	0,8%	3,2%	5,7%	-1,9%	2,3%
AG	0,432	0,453	0,355	0,253	0.198
Δ AG	21,7%	19,2%	26,8%	17%	12,5%
DT	0,358	0,376	0,278	0,215	0,175
Δ DT	0,8%	-0,011%	-0,7%	-0,5%	-0,6%
RF	0,358	0,380	0,282	0,216	0,176
Δ RF	0,8%	0,00%	0,7%	0%	0%

système. Nous avons aussi calculé la variation de la pertinence système, notée Δ, est calculée comme suit :

$$\Delta = \frac{(Pertinence\ avec\ expansion) - Baseline}{Baseline}$$

La Table 4 synthétise les résultats obtenus pour la collection SDA95. Les valeurs dans la ligne BE indiquent les résultats obtenus avec les requêtes étendue par la méthode aveugle. La ligne PRF reporte les résultats obtenus avec les requêtes étendues en utilisant la méthode de pseudo relevance feed back. Les valeurs dans la ligne AG sont obtenues pour les requêtes étendues en utilisant l'algorithme d'exploration basé sur les algorithmes génétiques. Les valeurs de la ligne DT correspondent aux expérimentations sur les requêtes étendues par apprentissage en appliquant le classifieur $J48$. Les valeurs de la ligne RF présentent les résultats obtenus pour les requêtes étendues par apprentissage en appliquant le classifieur $RandomForest$. Nous indiquons pour chacune de ces lignes les variations par rapport à la baseline.

Ces résultats montrent que les MAP obtenus avec les requêtes optimisées par les AG sont nettement meilleure en terme de MAP et de précision. En effet, ces résultats représentent une borne supérieur à atteindre idéalement avec l'expansion par apprentissage puisque l'expansion par AG explore l'espace des règles d'association tout en maximisant la MAP. L'expansion de requêtes par apprentissage apporte une amélioration de 0,8% des résultats de la recherche en terme de MAP par rapport à la *baseline*. Nous constatons d'autre part que notre approche est aussi performante aue PRF. L'expansion "aveugle" se basant sur les règles d'association s'est avérée inefficace pour la collection testée ce qui montre l'intérêt de l'approche de sélection des règles d'association que nous proposons dans ce travail. Ceci peut s'expliquer par la nature des instances d'apprentissages qui présentent un déséquilibre considérable entre le nombre d'instances négatives et positives. En effet, la classe positive est minoritaire (3% du

nombre total d'instances), ce qui ne permet pas au classifieur de mettre en évidence des plages de valeurs des attributs qui soit discriminantes pour cette classe comme le montrent les faibles valeurs des PVP dans les tables 2 et 3. Des techniques de sur-échantillonnage et sous-échantillonnage ont été utilisées pour remédier à ce problème de déséquilibre d'instances mais aucune de ces technique n'a permis d'améliorer les taux des PVP ni des résultats de la recherche.

6. Conclusion et travaux en cours

Nous avons présenté dans cet article, une approche d'expansion de requêtes par apprentissage qui se base sur les règles d'association entre termes. Le problème d'expansion est modélisé en tant qu'un problème de classification supervisée. Un processus exploratoire basé sur les algorithmes génétiques explore l'espace des règles d'association à la recherche des meilleurs termes d'expansion et génère en parallèle des instances d'entraînement qui sont utilisés par la suite pour construire un classifieur. La résolution du problème d'apprentissage se fait par arbre de décision et par la méthode $RandomForest$. Les expérimentations menées sur la collection de textes français SDA-95 montrent que les requêtes étendues par apprentissage pour les deux méthodes permettent d'améliorer les résultats de la recherche par rapport aux requêtes originelles et restent au même niveau que les résultats de PRF. Nous travaillons à présent sur une nouvelle modélisation du problème qui permettra de classer les règles d'association relatives à une requête selon leur capacité de fournir une expansion adéquate de la requête.

7. Remerciements

Ce travail est partiellement financé par le projet de collaboration Franco-Tunisien DGRST-CNRS :14/R 1401.

8. Bibliographie

Araujo L., Zaragoza H., Pérez-Agüera J. R., Pérez-Iglesias J., « Structure of morphologically expanded queries : A genetic algorithm approach », *Data Knowl. Eng.*, vol. 69, nº 3, p. 279-289, 2010.

Boughanem M., Tamine L., « A Study on Using Genetic Niching for Query Optimisation in Document Retrieval », *Advances in Information Retrieval, 24th BCS-IRSG European Colloquium on IR Research Glasgow, UK, March 25-27, 2002 Proceedings*, p. 135-149, 2002.

Bouziri A., Latiri C., Gaussier É., Belhareth Y., « Learning Query Expansion from Association Rules Between Terms », *KDIR 2015 - Proceedings of the International Conference on Knowledge Discovery and Information Retrieval, part of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2015), Volume 1, Lisbon, Portugal, November 12-14, 2015*, p. 525-530, 2015.

Buckley C., Salton G., Allan J., Singhal A., « Automatic Query Expansion Using SMART : TREC-3 », *Proceedings of the* 3^{rd} *Text REtrieval Conference*, 1994.

Carpineto C., Romano G., « A survey of automatic query expansion in information retrieval », *ACM Computing Surveys (CSUR)*, vol. 44, n° 1, p. 1, 2012.

Chifu A., Mothe J., « Expansion sélective de requêtes par apprentissage », *in* M. Moens, C. Viard-Gaudin, H. Zargayouna, O. R. Terrades (eds), *CORIA 2014 - Conférence en Recherche d'Infomations et Applications- 11th French Information Retrieval Conference. CIFED 2014 Colloque International Francophone sur l'Ecrit et le Document, Nancy, France, March 19-23, 2014.*, ARIA-GRCE, p. 257-272, 2014.

Grefenstette G., « Use of semantic context to produce term association lists for text retrieval », *Proceedings of the* 15^{th} *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'92*, ACM Press, Copenhagen, Denmark, p. 89-97, June, 1992.

Haddad H., Chevallet J. P., Bruandet M. F., « Relations between Terms Discovered by Association Rules », *Proceedings of the Workshop on Machine Learning and Textual Information Access in conjunction with the* 4^{th} *European Conference on Principles and Practices of Knowledge Discovery in Databases, PKDD 2000*, Lyon, France, September, 2000.

Jones K. S., Walker S., Robertson S. E., « A probabilistic model of information retrieval : development and comparative experiments », *Information Processing and Management*, vol. 36, n° 6, p. 779-840, 2000.

Kumaran G., Allan J., « Adapting information retrieval systems to user queries », *Information Processing and Management*, vol. 44, n° 6, p. 1838-1862, 2008.

Latiri C., Haddad H., Hamrouni T., « Towards An Effective Automatic Query Expansion Process Using An Association Rule Mining Approach », *Journal of Intelligent Information Systems*, vol. 39, n° 1, p. 209-247, 2012.

Quinlan J., « Simplifying decision trees », *International Journal of Man Machine Studies*, vol. , n° 27, p. 221-234, 1987.

Rijsbergen C. V., *Information Retrieval*, Butterworths, London, 1979.

Rungsawang A., Tangpong A., Laohawee P., Khampachua T., « Novel Query Expansion Technique Using Apriori Algorithm », *Proceedings of the* 8^{th} *Text REtrieval Conference, TREC 8*, Gaithersburg, Maryland, p. 453-456, November, 1999.

Ruthven I., Lalmas M., « A survey on the use of relevance feedback for information access systems », *Knowledge Engineering Review*, vol. 18, n° 2, p. 95-145, 2003.

Salton G., McGill M. J., *Introduction to Modern Information Retrieval*, McGraw-Hill, 1983.

Tangpong A., Rungsawang A., « Applying Association Rules Discovery in Query Expansion Process », *Proceedings of the* 4^{th} *World Multi-Conference on Systemics, Cybernetics and Informatics, SCI 2000*, Orlando, Florida, USA, July, 2000.

Zaki M. J., « Mining Non-Redundant Association Rules », *Data Mining and Knowledge Discovery*, vol. 9(3), p. 223-248, 2004.