Enrichissement de requêtes pour la recherche documentaire selon une classification non supervisée

Christian Raymond, Patrice Bellot, Marc El-Bèze

Enrichissement de requêtes pour la recherche documentaire selon une classification non supervisée

Query expansion for information retrieval by means of an unsupervised classification

Christian Raymond, Patrice Bellot, Marc El-Bèze

Laboratoire d'Informatique d'Avignon (LIA) - Université d'Avignon

Agroparc BP 1228 - 84911 Avignon Cedex 9 (France)
{christian.raymond,patrice.bellot,marc.elbeze}lia.univ-avignon.fr

Résumé

Une difficulté majeure dans l'utilisation d'un système de recherche documentaire est le choix du vocabulaire à employer pour exprimer une requête. L'enrichissement de la requête peut prendre plusieurs formes : ajout de mots extraits automatiquement des documents rapportés, réestimation des poids attribués à chacun des mots de la requête initiale, etc. *Le système de recherche documentaire SIAC est utilisé pour extraire un premier jeu de documents à partir d'une requête. Une méthode de classification non supervisée, à base d'arbres de décision, est ensuite exploitée pour classer les phrases des documents trouvés selon qu'elles contiennent ou non certains mots extraits automatiquement de l'ensemble des documents rapportés. À chaque nœud de l'arbre, peut être associée une expression booléenne mettant en jeu les mots sélectionnés lors de la classification. Nous montrons, à l'aide des données de la seconde campagne d'évaluation Amaryllis, que la réécriture de la requête suivant les expressions booléennes correspondant aux meilleures feuilles permet d'améliorer la précision de la recherche documentaire.*

Mots Clef

Recherche documentaire, enrichissement de requête, classification automatique, arbres de décision non supervisés.

Abstract

Natural language query formulation is a crucial task in the information retrieval (IR) process. Automatic expanding and refining of queries can be realized in different ways : extracting some words from top retrieved documents (retrieval feedback) or from thesauri, computing new query term weights according to top retrieved documents... In this paper, the information retrieval system SIAC is employed to obtain an initial set of documents from a query. Then, a classification method employing unsupervised decision trees (UDTs) is performed to classify the document retrieved sentences according to some words extracted automatically from these documents (some sentences contain the chosen words, some do not). A boolean expression composed of these selected words is directly associated to each decision tree node. This paper shows that expanding queries with the words connected with the best nodes allows to significantly improve retrieval precision.

Keywords

Information retrieval, query expansion, automatic classification, unsupervised decision trees.

1 Introduction

Il arrive souvent qu'un utilisateur d'un système de recherche documentaire exprime le contenu conceptuel de l'information requise avec des mots qui ne correspondent pas aux mots apparaissant dans les documents pertinents. Ce problème de vocabulaire est encore plus important dans le cas de requêtes courtes. De nombreux efforts ont donc été entrepris sur le développement de méthodes d'enrichissement de requêtes. Les méthodes les plus courantes extraient automatiquement les mots dans les documents retournés en tête de liste. Elles sont connues sous l'appellation de *retrieval feedback* ou *pseudorelevance feedback* [4], [15], [16]. D'autres techniques s'appuient sur des thesaurus construits manuellement ou automatiquement

pour guider l'utilisateur vers une reformulation de sa requête à travers un processus interactif [6], [8]. Pour une description détaillée de ces méthodes, voir par exemple [5], [7].
Cet article présente l'exploitation d'une méthode originale non supervisée de classification de textes pour l'enrichissement de requêtes. Dans la première section, les systèmes de recherche documentaire et la méthode de classification sont présentés . La deuxième section, présente les hypothèses et propose les modèles qui seront évalués puis analysés dans la section 3.

1.1 Le système SIAC

Un système de recherche documentaire permet à un utilisateur qui soumet une requête de rechercher et retrouver les documents jugés pertinents par rapport à cette requête. Ces réponses sont ordonnées en fonction du calcul d'une similarité entre les documents et la requête. La plupart des systèmes se basent sur les mots communs à la requête et aux documents pour effectuer leurs recherches.
Le système de recherche documentaire SIAC (décrit dans [2]) du LIA a été originellement développé afin de tester des méthodes de classification et de segmentation automatiques non supervisées. Il intègre un moteur de recherche et d'indexation basé sur le modèle vectoriel. SIAC a participé aux campagnes internationales d'évaluation Amaryllis (1997 et 1999) et TREC-7. Il est prévu de l'utiliser aussi pour la piste "Questions/Réponses" de TREC.
Les systèmes vectoriels basés sur les propositions de Salton [13] fonctionnent de la manière suivante : dans un espace à n dimensions constitué de D_i documents identifiés par un ou plusieurs termes d'indexation T_j (éventuellement pondérés en fonction de leur importance dans le document, voir par exemple [9], [11], [14]) la requête de l'utilisateur est modélisée sous la forme d'un vecteur $(q_1, q_2, ..., q_t)$ où q_i représente le poids du i^e terme et on calcule chaque similarité $s_j(Q, D_j)$, où j représente le j^e document ($j = 1; 2; ...; n$). Plusieurs mesures de similarité (telles le produit scalaire ou le cosinus) permettent de classer les documents retrouvés en fonction de leur pertinence par rapport à la requête. Pour en savoir plus, voir [1].
La méthode de pondération des termes TF.IDF (*Term Frequency, Inverse Document Frequency*) utilisée dans le moteur SIAC, tient compte du nombre d'apparitions du terme dans le document (ou la requête) et du nombre de documents qui contiennent ce terme dans le corpus. Elle est définie de la manière suivante :

$$\begin{aligned} TFIDF(w,d) &= TF_{w,d}.IDF_{w,d} \\ &= TF_{w,d}.((\log_2 \frac{N}{DF_w}) + 1) \quad (1) \end{aligned}$$

avec : w un terme, d un document ou une requête, $TF_{w,d}$ le nombre d'apparitions de w dans d, DF_w le nombre de documents du corpus qui contiennent au moins une fois w et N le nombre total de documents dans le corpus.

1.2 Une méthode de classification non supervisée

SIAC permet de classer les documents qui ont été retournés à partir de la requête. Cette classification est faite à l'aide d'un arbre de décision (*Semantic Classification Tree*, voir [10]) non-supervisé. Pour chaque requête, un arbre est crée. Il permet d'obtenir des feuilles contenant des individus thématiquement proches. Appliquée aux phrases des documents, cette classification permet de trouver des frontières thématiques à l'intérieur d'un document, lorsque deux phrases d'un même document se retrouvent dans des feuilles différentes.

Arbre de décision. Pour construire un tel arbre, on doit définir [3] :
- un ensemble de questions à poser aux individus (ici les phrases des documents trouvés) ; nous choisissons des questions telles que chaque individu peut y répondre par l'affirmative ou la négative ; suivant sa réponse, l'individu est transféré dans le nœud fils correspondant à la réponse "OUI" ou dans le fils correspondant à la réponse "NON"
- une règle pour déterminer les questions à poser aux individus ;
- un critère d'arrêt déterminant l'ensemble des feuilles de l'arbre.

Questions posées aux individus. À chaque nœud de l'arbre est associée une question de la forme "*les individus contiennent-ils le terme x ?*". Cette question permet de subdiviser le nœud courant en deux nœuds fils qui comprennent respectivement les individus du nœud courant qui contiennent et qui ne contiennent pas le terme x. Dans notre cas, x désigne n'importe quel terme présent dans l'ensemble des documents à classer. À chaque nœud de l'arbre, il faut calculer quel est le terme x qui conduit à la meilleure répartition en deux sous-partitions des individus de ce nœud.

Le critère de sélection des questions. Le critère de sélection d'une question est souvent fonction du gain en entropie observé avant et après l'affectation des individus dans de nouvelles partitions suivant la question considérée (formule (2)). L'entropie d'une partition S composée des classes[1] $c_1, c_2, ..., c_k$ de probabilités respectives $p_1, p_2, ..., p_k$ est définie par :

$$Entropie : H_S = \sum_{i=1}^{k} -p_i . \log_2 p_i \quad (2)$$

[1]Dans notre cas, k=2, la classe des documents pertinents et celle des documents non pertinents

Les probabilités utilisées dans le calcul de l'entropie sont celles que la requête soit générée à partir des textes du nœud considéré. Cette probabilité[2] p peut être calculée simplement en fonction d'un modèle unigramme (voir formule 3).
Soient r la requête, w_j le j^e terme de r, S_n un nœud de l'arbre, I_i un individu et $Z(w_j, I_i)$ le nombre d'apparitions de w_j dans les individus de S_n. Les signes $||$ sont utilisés pour représenter la taille d'un individu (le nombre de termes (occurrences) qu'il contient).

$$p(r = w_1, ..., w_n \mid \bigcup_{i/I_i \in S_n} I_i) = \prod_{j=1}^{n} p(w_j \mid \bigcup_{i/I_i \in S_n} I_i) = \frac{\prod_j (\sum_i Z(w_{j,I_i}))}{|\bigcup_{i/I_i \in S_n} I_i|^n} \quad (3)$$

Le critère d'arrêt, devant être défini pour éviter d'obtenir autant de feuilles que d'individus à classer, est choisi dans nos expériences comme étant une valeur seuil empirique du gain minimal en entropie autorisé pour subdiviser un nœud.

2 Classification et enrichissement

2.1 Exploitation d'une classification pour l'enrichissement

SIAC permet de regrouper dans les feuilles de l'arbre, en fonction de la requête, les phrases issues des documents retournés, "thématiquement proches" les unes des autres. Les questions posées à chaque nœud de l'arbre étant de la forme : "*les individus contiennent-ils le mot X ?*" il est possible de représenter les individus d'une feuille de l'arbre par une expression booléenne. L'expression booléenne représentant une *bonne* feuille pourrait se révéler efficace pour enrichir la requête initiale (voir requêtes 24 et 6, respectivement dans les tableaux 1 et 2). En effet, si l'arbre a réussi à construire une feuille possédant au final une grande proportion de documents pertinents, nous pouvons espérer d'une part que l'exploitation de cette expression, nous permette de positionner ces documents pertinents en tête de liste, d'autre part que les mots trouvés par l'arbre nous permettent de trouver de nouveaux documents pertinents. Toutefois, deux types de mots sont présents dans ces expressions booléennes. Les mots pour lesquels les individus ont répondu NON et ceux pour lesquels ils ont répondu OUI. Les premiers sont des mots que l'on ne souhaite pas retrouver dans les documents (les mots "négatifs"), les seconds sont des mots que l'on souhaite retrouver dans les documents (les mots "positifs"). Les modèles vectoriels de recherche documentaire ne prenant pas en compte naturellement les mots "négatifs", seuls les mots "positifs" seront utilisés pour l'enrichissement dans cet article.
Dans un premier temps, il faut trouver l'expression booléenne la plus utile ou *performante*, ce qui amène à réfléchir sur ce qu'est une *bonne* feuille.
Et pour finir, une pondération pour les mots issus de l'expression booléenne doit être calculée. En effet, certains mots contribuent plus que d'autres au partitionnement donnant la feuille qui nous intéresse. D'autre part, ces mots peuvent être des mots existants dans la requête originelle ou bien des mots nouveaux. Par exemple la meilleure feuille (F_6) (arbre n° 1 de la figure 1) peut être représentée par l'expression[3] :
(NON(PUDEUR)) ET (NON(CAIRE)) ET (TRADE) ET (ATTENTAT)

attentat conspiration terroriste violence urbain secte de+le Davidiens Emeutes sanglant complot terroriste ***attentat trade***

TAB. 1 – Requête n° 24 OT1 (les mots en gras sont issus de l'enrichissement)

dumping social grêve représentant personnel acquis social pouvoir argent exploitation profit ***préavis faim***

TAB. 2 – Requête n° 6 OT2 (les mots en gras sont issus de l'enrichissement)

[2]Un problème se pose dans ce calcul de probabilité. Lorsqu'un mot de la requête ne se retrouve dans aucun document la probabilité associée à ce mot est 0 et le résultat final se retrouve lui même à zéro. Le problème a été partiellement contourné en considérant qu'une occurrence de chaque mot de la requête est présente dans chaque phrase de la feuille.

[3]La requête est à situer dans un contexte antérieur aux attentats du 11 septembre 2001, la première version de cet article ayant été écrit en juin 2001. L'enrichissement fait référence ici à l'attentat du world trade center de 1993 et son actualité vis à vis des derniers événements n'est que pure coincidence.

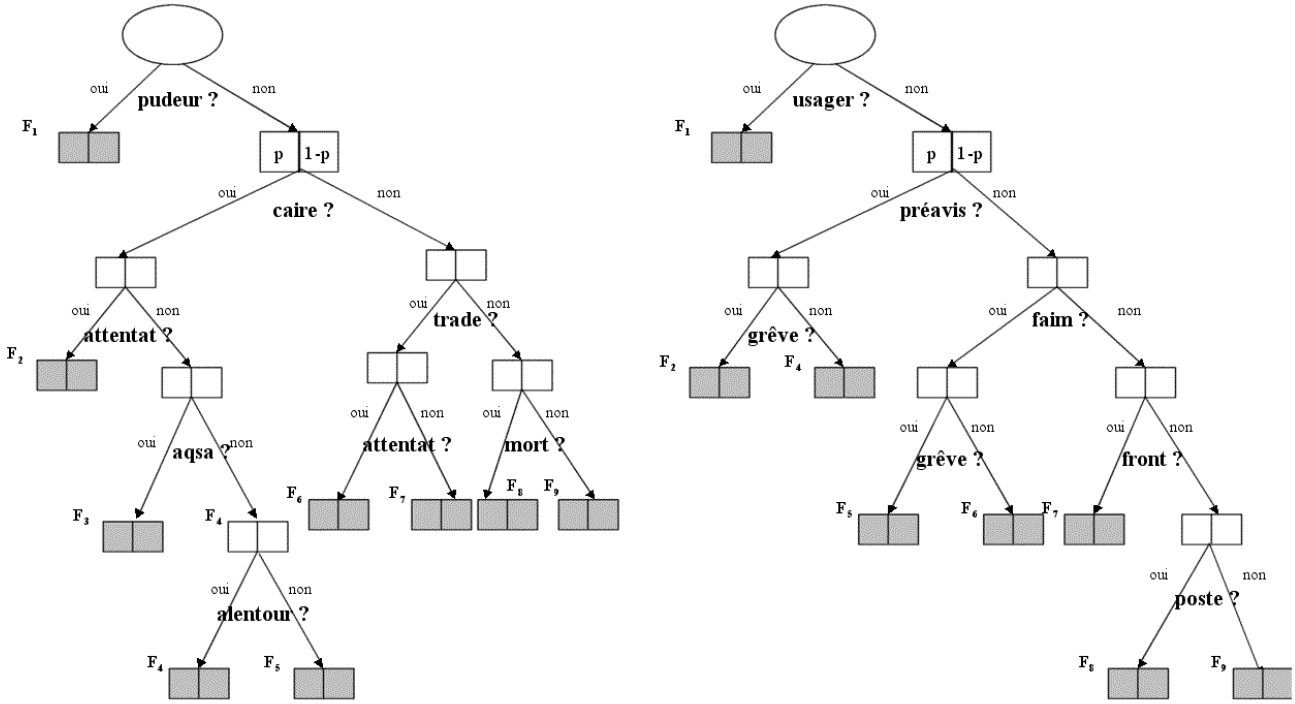

FIG. 1 – Arbres générés à partir des documents retrouvés pour les requêtes n° 24(OT1) et 6(OT2). (Amaryllis'99 - Corpus OFIL OD1)

2.2 Recherche de la meilleure feuille

L'objectif premier est de repositionner les documents pertinents en tête des réponses (amélioration de la précision sur les premiers documents) plutôt que d'augmenter leur nombre (rappel). Les meilleures feuilles doivent être sélectionnées en fonction du critère lié à cet objectif : la précision. Dans les expériences décrites dans cet article, nous utilisons la liste des documents pertinents donnée par les organisateurs de la campagne Amaryllis (*cf* section 3.1). Ce choix expérimental induit un biais et limite un peu la portée de l'étude, mais il permet de ne pas avoir à répondre immédiatement au difficile problème de la détection de la meilleure feuille de l'arbre. Il permet ainsi de valider la méthode d'enrichissement en s'affranchissant temporairement de ce problème.
Il est possible d'utiliser la précision globale d'une feuille :

$$\frac{nombre\ de\ documents\ pertinents\ dans\ la\ feuille}{nombre\ total\ de\ documents\ dans\ la\ feuille} \quad (4)$$

ou la précision sur les premiers documents. La précision globale privilégie trop les feuilles peu peuplées. En effet, une feuille ne contenant qu'un document mais pertinent sera considérée meilleure qu'une feuille contenant 20 documents dont 19 sont pertinents. Il est pourtant légitime de penser que l'expression booléenne permettant d'accéder à cette feuille va influer sur un plus grand nombre de documents et être plus efficace dans l'opération de repositionnement des documents pertinents en tête de liste, notamment lorsque la feuille choisie possède une précision sur les premiers documents supérieure à celle de la liste de documents avant classification. Il est vrai qu'inversement, une feuille contenant un seul document qui est en outre pertinent aura une moins bonne précision à 5 documents qu'une feuille contenant 2 documents pertinents parmi les 5 premiers puis 200 documents non pertinents ensuite. La sélection de la meilleure feuille est donc un problème ouvert même lorsqu'on connaît la liste des documents à trouver.

2.3 Pondération des mots

La pondération des mots tirés de l'arbre est calculée comme les autres mots de la requête originelle selon le critère TF.IDF utilisé dans SIAC. Les pondérations des mots de la nouvelle requête sont données par la formule suivante :

$$Q_{new}(w) = \alpha.TF.IDF(w, Q_{new}) \quad (5)$$

Où α vaut 1 si w est un mot de la requête originelle, et sinon une valeur déterminée de façon empirique dans le cas ou w est un mot de l'expression booléenne.

3 Évaluation

Les évaluations de notre méthode d'enrichissement se font sur les données des campagnes Amaryllis. Avant d'analyser nos résultats (section 3.2), la campagne Amaryllis ainsi que les critères d'évaluations utilisés dans cet article sont présentés dans la section 3.1.

3.1 Critères et contexte d'évaluation

Campagnes d'évaluation Amaryllis. Afin de pouvoir comparer systèmes et méthodes, des campagnes d'évaluation sont organisées depuis quelques années. C'est le cas des campagnes francophones Amaryllis (voir notamment [12]) dont le premier cycle s'est déroulé de 1996 à 1997 et le second de 1998 à 1999. Des thèmes sont fournis aux participants. Ces thèmes contiennent différents champs textuels à partir desquels doivent être construites, automatiquement ou manuellement, les requêtes (voir exemple dans le tableau 3).
Les évaluations décrites dans cet article utilisent les deux jeux de thèmes OT1 et OT2 (chacun contient 26 thèmes) sur les corpus OD1 et OD2 de l'OFIL. Les caractéristiques des corpus sont disponibles dans le tableau 4.

Évaluation en recherche documentaire. L'évaluation d'un système de recherche documentaire peut se faire selon les critères de précision et de rappel définis ci-dessous :

La précision La précision d'une liste de documents est la proportion de documents pertinents dans cette liste. Elle est définie par :

$$précision = \frac{nombre\ de\ documents\ pertinents\ trouvés}{nombre\ de\ documents\ trouvés} \quad (6)$$

Le rappel Le rappel rend compte de la quantité de documents pertinents rapportés par rapport au nombre de documents pertinents dans le corpus. Autrement dit, le rappel est le taux de documents pertinents trouvés par rapport au nombre de documents pertinents à trouver. Le rappel d'une liste est défini par :

$$rappel = \frac{nombre\ de\ documents\ pertinents\ trouvés}{nombre\ de\ documents\ pertinents\ à\ trouver} \quad (7)$$

Courbe rappel/précision Une des manières de tenir compte à la fois du rappel et de la précision d'un système est d'interpoler les valeurs de précision correspondant à différents niveaux de rappel. Les niveaux standards de rappel varient entre 0 et 1 suivant un pas de 10%. La règle d'interpolation utilisée pour les campagnes TREC et Amaryllis définit la précision pour un niveau de rappel i comme étant la valeur maximale de précision pour tout niveau de rappel supérieur ou égal à i. Suivant cette définition une valeur de précision pour un rappel nul correspond au niveau maximal de précision obtenu pour un rappel quelconque.

domaine : *International*
sujet : *La séparation de la Tchécoslovaquie*
question : *Pourquoi et comment avoir divisé la Tchécoslovaquie et quelles ont été les répercussions économiques et sociales ?*
compléments : *Prendre en compte les différentes versions présentées*
concepts : *Partition de la Tchécoslovaquie, causes et modalités de la partition, création de la Slovaquie et de la république Tchèque, points de vue, économie.*

TAB. 3 – Thème n° 1 de la campagne Amaryllis'99

Titre de la collection	Nature	Taille en Mo	Nombre de documents	Nombre total de mots (en millions)	Nombre de lemmes différents (en milliers)
OFIL OD1	Articles Journalistiques	33	11 016	2.4	53.5
OFIL OD2		34	9 287	3.2	60

TAB. 4 – Les corpus OD1 et OD2

3.2 Classification des phrases des documents et enrichissement

L'exploitation des mots de l'arbre pour l'enrichissement des requêtes a été testée avec les mots pour lesquels les phrases contenues dans la meilleure feuille ont répondu OUI. Le nombre de requêtes pouvant être enrichies avec des mots "positifs" est de 8 requêtes sur 26 pour le corpus OD1, 7 pour le corpus OD2. La feuille contenant les phrases ayant répondu NON à toutes les questions est la meilleure feuille pour les autres requêtes. Pour celles-ci, aucun enrichissement n'a donc été effectué.

Réordonnancement des documents. Les résultats suivants sont les résultats d'une nouvelle recherche avec les nouvelles requêtes enrichies avec les mots en questions. Le poids des mots est calculé comme décrit dans la formule 5 avec $\alpha = 1$. L'objectif premier étant d'améliorer la précision, sans toutefois perdre en rappel, la nouvelle recherche est effectuée sur un nouveau corpus constitué des documents retournés avec la première recherche (lignes notées "Ajout" dans les tableaux. Ceci a pour but de réordonner les documents. Nous constatons une amélioration absolue de la précision sur les 30 premiers documents de l'ordre de 10% avec $\alpha = 1$, de 14% avec $\alpha = 6$ (figure 2). Toutefois le repositionnement des documents avec la requête originelle fait chuter les performances. En effet les mots de la requête sont maintenant plus représentés dans le corpus constitué des documents rapportés par la première recherche, ce qui a pour effet de modifier le critère IDF et d'influer sur l'ordre des documents et généralement le rendre moins performant. Dans le cas d'un réordonnancement des documents les valeurs IDF du corpus initial semblent devoir être conservées.

Nouvelle recherche. Le tableau 6 et la figure 3 montrent les résultats des précédentes expériences sur le corpus complet. Ces expériences ont donc des incidences sur le rappel. Les requêtes étant différentes, la liste retournée par le moteur de recherche en est changée. Des améliorations du même ordre sont constatées (figure 3) avec une hausse absolue de la précision sur les 30 premiers documents de 6% avec $\alpha = 1$ et 8% avec $\alpha = 3$. Le tableau 7 montre les résultats sur l'ensemble des 26 requêtes dont 18 n'ont pas été modifiées.

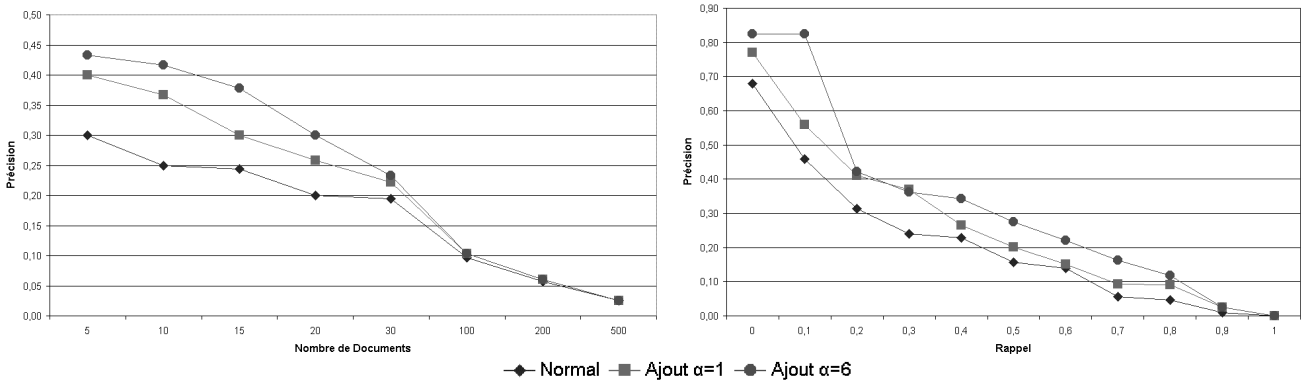

FIG. 2 – Évaluation du réordonnancement sur 7 requêtes enrichies OT1 - extraits Corpus OD2 - Amaryllis'99)

Expériences	Précision pour un rappel 0.00	Précision pour un rappel 0.10	Précision des 5 premiers documents	Précision des 10 premiers documents	Précision des 15 premiers documents	Nombre de documents pertinents
Référence OD1	0.43	0.37	0.23	0.28	0.31	100
Ajout ($\alpha = 1$) extraits OD1	0.45 (+2%)	0.30 (-7%)	0.27 (**+4%**)	0.32 (**+4%**)	0.29 (-2%)	100
Ajout ($\alpha = 4$) extraits OD1	0.54 (**+11%**)	0.35 (-2%)	0.40 (**+17%**)	0.33 (**+5%**)	0.31 (+0%)	100

TAB. 5 – Résultats du réordonnancement des documents sur les 7 requêtes enrichies OT1 -Corpus OD1 Amaryllis'99)

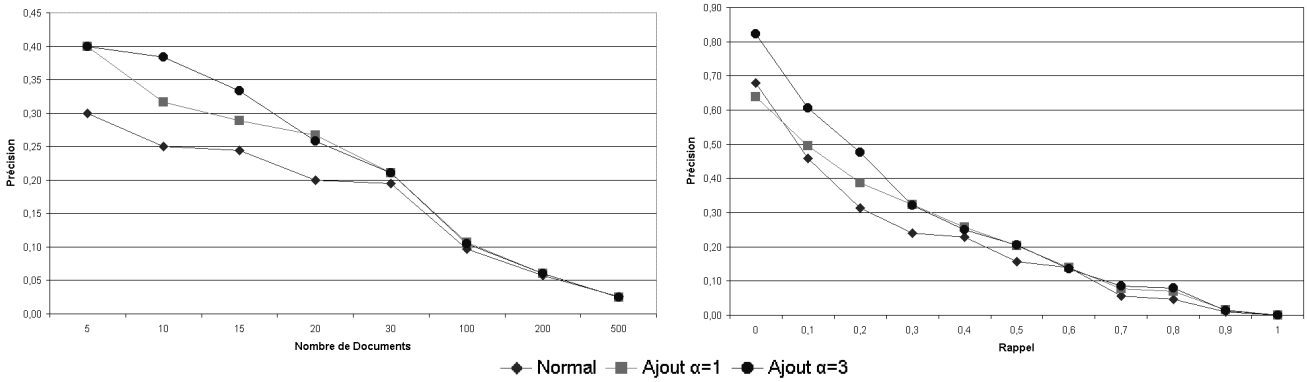

FIG. 3 – Évaluation de la nouvelle recherche sur 7 requêtes enrichies OT1 - Corpus OD2 - Amaryllis'99.

Expériences	Précision pour un rappel 0.00	Précision pour un rappel 0.10	Précision des 5 premiers documents	Précision des 10 premiers documents	Précision des 15 premiers documents	Nombre de documents pertinents
Référence OD1	0.43	0.37	0.23	0.28	0.31	100
Ajout ($\alpha = 1$) OD1	0.47 (+3%)	0.39 (+2%)	0.30 (**+7%**)	0.35 (**+7%**)	0.34 (+3%)	98

TAB. 6 – Résultats de la nouvelle recherche sur les 7 requêtes enrichies OT1 -Corpus OD1 Amaryllis'99.

Expériences	Précision pour un rappel 0.00	Précision pour un rappel 0.10	Précision des 5 premiers documents	Précision des 10 premiers documents	Précision des 15 premiers documents	Nombre de documents pertinents
Référence OD2	0.62	0.44	0.35	0.29	0.27	362
Ajout ($\alpha = 1$) OD2	0.61 (-1%)	0.45 (+1%)	0.37 (+2%)	0.30 (+1%)	0.28 (+1%)	359

TAB. 7 – Résultats de la nouvelle recherche avec la totalité des 26 requêtes OT1 Corpus OD2

4 Conclusion et perspectives

Les résultats obtenus montrent que l'exploitation des mots tirés de l'expression booléenne conduisant à la meilleure feuille augmente la précision. L'exploitation des mots "positifs" a entraîné des améliorations significatives absolues de la précision de l'ordre de 10% (ce qui correspond à des améliorations relatives de la précision de l'ordre de 30%). L'enrichissement au moyen des mots de l'expression booléenne nécessite une étude plus approfondie portant notamment sur les perspectives de développement et d'améliorations suivantes. Il faut développer une méthode d'accès automatique à la meilleure feuille. Le chemin conduisant à la meilleure feuille pourrait être trouvé selon certaines valeurs numériques associées aux arbres de décision, comme la valeur du critère utilisé dans le choix de la meilleure question à poser, la probabilité associée à chaque feuille, ou (et) des valeurs facilement calculables *a posteriori* comme la population de chacune des feuilles, leur profondeur dans l'arbre, *etc.*

L'utilisateur pourrait vouloir préciser certaines notions qu'il ne veut pas retrouver, par exemple "*Je désire des textes au sujet de la réaction de fusion atomique, le phénomène de fission ne m'intéresse pas*". Il serait alors particulièrement intéressant d'introduire dans la requête des mots exprimant un point de vue "négatif" nuancé (car des documents parlant à la fois de fusion et de fission peuvent être intéressants) afin de rejeter plus ou moins fortement certains documents (dans l'exemple le mot "fission"). L'expression booléenne représentant la meilleure feuille de l'arbre nous donne des mots, ceux précédés de l'opérateur NON, exprimant ce point de vue négatif. L'adaptation du modèle vectoriel pour la prise en compte de la négation constitue l'une des perspectives les plus intéressantes de l'étude présentée dans ce papier.

Références

[1] R. Baeza-Yates, B. Ribeiro-Neito. *Modern information retrieval.* ACM press books, Addison-Wesley edition, 1999.

[2] Patrice Bellot. *Méthodes de classification et segmentation locales non supervisées pour la recherche documentaire.* Thèse de doctorat, Université d'Avignon et des pays du Vaucluse, janvier 2000.

[3] L. Breiman, J. Friedman, R. Olshen, C. Stone. *Classification and Regression Trees.* Wadsworth, Belmont, CA, 1984.

[4] Chris Buckley, Gerard Salton, James Allan, Amit Singhal. Automatic query expansion using SMART : TREC-3, NIST special publication. In *Text REtrieval Conference*, pages 69–80, 1995.

[5] Claudio Carpineto, Renato de Mori, Giovanni Romano, Brigitte Bigi. An information-theoretic approach to automatic query expansion. 19(1) : 1–27, 2001.

[6] James W. Cooper, Roy J. Byrd. Lexical navigation : Visually prompted query expansion and refinement. In *ACM DL*, pages 237–246, 1997.

[7] Claude De Loupy. *Évaluation de l'Apport de Connaissances Linguistiques en Désambihuisation Sémantique et Recherche Documentaire.* Thèse de doctorat, Université d'Avignon et des pays du Vaucluse, novembre 2000.

[8] Gregory Grefenstette. *Explorations in automatic thesaurus discovery.* Kluwer Academic Publishers, Dordrecht, NL, 1994.

[9] D.K. Harman. *Ranking algorithms*, in *Information Retrieval*, édité par W. B. Frakes et R. Baeza-Yates. Englewood Cliffs, NJ : Prentice Hall : 363–392.

[10] R. Kuhn, R. De Mori. *The Application of Semantic Classification Trees to Natural Language Understanding*, volume 17, chapitre 5, pages 449–460. IEEE Transactions on Pattern Analysis and Machine Intelligence, Mai 1995.

[11] K.L. Kwok. A new method of weighting query terms for ad-hoc retrevial. *Actes de ACM/SIGIR'96 Conference on Research and Development in Information retrevial, Zurich Suisse*, pages 187–195, 1996.

[12] K. Lespinasse, P. Kremer, D. Schibler, L. Schmitt. *Evaluation des outils d'accès à l'information textuelle, les expériences américaine (Trec) et française (Amaryllis)*, Langues, John Libbey, volume 2, n° 2, pages 100–109, 1999.

[13] G. Salton. *The SMART retrieval system.* Pre Hall, Englewood Cliffs NJ, USA, 1971.

[14] G. Salton, J. Allan. Automatic text decomposition and structuring. *Actes de RIAO'94*, pages 6–29, 1994.

[15] Gerard Salton, Christopher Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5) : 513–523, 1988.

[16] Jinxi Xu, W. Bruce Croft. Query expansion using local and global document analysis. In *Research and Development in Information Retrieval*, pages 4–11, 1996.