Représentation Dynamique de Documents pour une Recherche Documentaire Intelligente

Rachid Arezki* — **Pascal Poncelet*** — **Gérard Dray*** — **David william Pearson****

* *Centre LGI2P EMA, Site EERIE Parc Scientifique Georges Besse*
30035 Nimes Cedex 1
{rachid.arezki, pascal.poncelet, gerard.dray }@ema.fr

** *EURISE, Université Jean Monnet de Saint-Etienne*
23 rue du Docteur Michelon, 42023 Saint-Etienne
david.pearson@univ-st-etienne.fr

RÉSUMÉ. *Avec le développement des supports de stockage, la quantité de documents disponibles ne cesse de croître. Il est donc nécessaire de disposer d'un système de recherche documentaire capable d'appréhender de manière efficace ces quantités énormes de documents. Dans cet article, nous proposons DIIR (Dynamic Indexing for Information Retrieval), un modèle vectoriel de recherche documentaire où la représentation des documents se fait de manière dynamique afin d'enrichir la requête et d'adapter les poids des termes dans les documents du corpus. L'étude empirique effectuée sur des données provenant de la campagne d'évaluation AMARYLLISS confirme la pertinence de notre approche.*

ABSTRACT. *With the development of internet and storage devices, the quantity of available documents increases quickly. It is necessary to have an information retrieval system able to apprehend efficiently these enormous quantities of documents. In this paper we propose DIIR (Dynamic Indexing for Information Retrieval), an information retrieval model based on the vector space model where the representation of documents is made dynamically in order to expand requests and to adapt the term weights of documents. An evaluation of the model on real data from the evaluation campaign AMARYLLISS confirms the relevance of our approach.*

MOTS-CLÉS : *Indexation Dynamique, Recherche Documentaire, Modèle Vectoriel, Analyse du Contexte Global.*

KEYWORDS: *Dynamic Indexing, Information Retrieval, Vector Space Model, Global Context Analysis.*

1. Introduction

Avec le développement des supports de stockage, la quantité de documents disponibles ne cesse de croître. Il est donc nécessaire de disposer d'un système de recherche documentaire capable d'appréhender de manière plus efficace ces quantités énormes de documents. Un des problèmes majeur de la recherche d'information est la formulation des requêtes. Blair et Maron [BLA 85] avaient montré que la faible performance des systèmes de recherche d'information est dû à l'incapacité des utilisateurs de formuler les requêtes adéquates. En effet, la requête initiale de l'utilisateur est souvent exprimés par une liste de termes souvent très réduite qui exprime mal les besoins en information de l'utilisateur. Pour remédier à ce problème, une solution consiste à enrichir automatiquement la requête afin d'améliorer la qualité des documents retrouvés.

Les techniques d'enrichissement automatique de requêtes peuvent être classées en deux catégories : celles basées sur l'analyse du contexte local et celles basées sur l'analyse du contexte global.

Les techniques basées sur l'analyse du contexte local permettent d'identifier les relations entre termes afin d'enrichir les requêtes et cela par l'analyse des documents retrouvés (les mieux classés) [ATT 77, BUC 94, CRO 97]. Dans [BUC 94], les auteurs proposent une technique qui suppose que les premiers documents retrouvés (les mieux classés) sont pertinents, ensuite la requête est enrichie suivant la méthode standard de relevance feedback [ROC 71]. Une méthode similaire est utilisée dans [CRO 97], où les premiers documents retrouvés sont utilisés pour re-estimer les probabilités des termes.

Les techniques basées sur l'analyse du contexte global permettent d'identifier les relation entre termes et documents par l'analyse du corpus documentaire. Une des techniques d'analyse du contexte globale est la classification des termes [JON 70], où les termes d'indexation sont regroupés par classes, en se basant sur leurs co-occurrences. Ensuite ces classes de termes sont utilisées pour l'enrichissement des requêtes. Des méthodes d'enrichissement de requête par un thésaurus de similarité ont été proposés [QIU 93, JIN 94]. Un thésaurus de similarité est une matrice terme-terme construite à partir du corpus documentaire, où chaque terme est représenté par un vecteur de documents dans un espace de documents. Ainsi dans [QIU 93] une méthode d'enrichissement de requête par un thésaurus de similarité a été proposée, l'approche consiste à rajouter à la requête des termes issus du thésaurus. Le choix des termes à rajouter se fait par un calcul de similarité entre les termes de la requête et ceux du thésaurus.

Le modèle LSI (Latent Semantic Indexing) [DEE 90, LAN 98] est aussi considéré comme une méthode d'analyse du contexte global [XU 00]. LSI utilise une décomposition en valeurs singulières sur la matrice représentative du corpus documentaire (*documents* $\times$ *unités linguistiques*), cette décomposition permet d'extraire les principales associations entres les unités linguistique d'un document. Enfin, dans [XU 00] une méthode combinant analyse local et global a été proposée.

Dans ce papier, nous proposons *DIIR* (Dynamic Indexing for Information Retrieval), un modèle vectoriel de recherche documentaire où la représentation des documents se fait de manière dynamique afin d'enrichir la requête et d'adapter les poids des termes dans les documents du corpus. En effet, *DIIR* tient compte de la requête utilisateur dans le processus d'indexation (choix des termes d'indexation et de leurs pondérations) et cela par l'analyse des corrélations entres les termes du corpus. A partir des termes sélectionnés dynamiquement il enrichit la requête et adapte les poids des termes des documents.

L'article est organisé de la manière suivante. Dans la Section 2, nous rappelons les principes de base de la recherche documentaire dans le cadre du modèle vectoriel standard. La Section 3 décrit le modèle de recherche documentaire proposé. Nous présentons dans la Section 4 une série d'expériences effectuées sur des données réelles provenant de la campagne AMARYLLISS. Enfin, dans la Section 5, nous concluons en résumant les avantages de notre approche et présentons les perspectives associées.

2. Recherche documentaire basée sur le modèle vectoriel standard

2.1. *Définition*

Dans le modèle vectoriel standard, chaque document d est représenté par un vecteur à n dimensions $(w_1, .., w_n)$, où w_i est le poids du terme t_i dans le document d. Un terme peut être un mot, un lemme où un composant (plusieurs mots ou lemmes ou stems). Cette représentation requiert la définition de l'ensemble des termes d'indexation et une méthode de pondération des termes. Pour chaque paire de documents (u, v) (où $\overrightarrow{u}$ et $\overrightarrow{v}$ sont leurs représentation vectorielle dans l'espace à n dimensions), une fonction de similarité $s(\overrightarrow{u}, \overrightarrow{v})$ doit être définie. Pour une requête de recherche donnée q (une requête est également du texte et peut être convertie en un vecteur $\overrightarrow{q}$ dans le même espace vectoriel que les autres documents), la recherche documentaire est effectuée par le calcul de similarité entre les documents et la requête. Ainsi, les documents les plus similaires à la requête sont proposés à l'utilisateur.
Plus formellement, la recherche documentaire basée sur le modèle vectoriel standard peut être définie par le 5-uplet $< X, Q, T, s, f >$, où X représente le corpus documentaire, Q est l'ensemble des requêtes, T représente l'ensemble des termes d'indexation, s est la fonction de similarité et f est la fonction de construction des termes d'indexation tel que $T = f(X)$.

2.2. *Construction de l'ensemble des termes d'indexation*

L'ensemble T des termes d'indexation est construit à partir de l'analyse des documents du corpus X. Les éléments de T sont sélectionnés de manière à être les plus discriminants. Il existe différentes manières de choisir les termes d'indexation [LUH 58, SAL 83, RIJ 79]. Le critère de sélection des termes le plus utilisé est la fréquence en documents (*document frequency*).

2.3. *Pondération des termes*

Le poids d'un terme représente le degré de son importance dans le document. Il y a principalement 3 facteurs permettant la pondération des termes :
(1) la fréquence dans le document,
(2) la fréquence en documents,
(3) la normalisation.
TF-IDF est la méthode de pondération qui a été la plus étudiée en recherche documentaire, où l'importance d'un terme est proportionnelle à la fréquence d'apparition de ce terme dans le document et inversement proportionnelle à la fréquence en documents (nombre de documents où le terme apparaît) [SAL 83]. Plus précisément, soit TF_i la fréquence d'apparition du terme t_i dans le document d, et soit DF_i la fréquence en document du terme t_i. Le poids du terme t_i dans le document d, noté w_i, est calculé comme suit :

$$w_i = \frac{TF_i}{\sum_j TF_j} \times log(N/DF_i)$$

où N représente le nombre de documents dans le corpus.

2.4. *Mesures de similarité*

Il existe différentes mesures permettant de calculer la similarité entre deux documents. La mesure de similarité la plus utilisée est le cosinus de l'angle entre les vecteurs représentant les documents. Soit $\overrightarrow{d_i}$, $\overrightarrow{d_j}$ deux vecteurs de documents, la similarité du *cosinus* entre ces deux documents est définie par :

$$sim(\overrightarrow{d_i}, \overrightarrow{d_j}) = \frac{\overrightarrow{d_i} \bullet \overrightarrow{d_j}}{\mid \overrightarrow{d_i} \mid \times \mid \overrightarrow{d_j} \mid}$$

où $\mid \overrightarrow{d_i} \mid$ représente la norme euclidienne du vecteur $\overrightarrow{d_i}$ et $\bullet$ représente le produit scalaire entre $\overrightarrow{d_i}$ et $\overrightarrow{d_j}$.

3. Le Modèle DIIR

L'objectif d'un système de recherche documentaire est de répondre par des documents pertinents à une requête souvent exprimée en langage naturel. Afin de répondre de manière plus adaptée, dans *DIIR* les termes d'indexation sont sélectionnés de manière dynamique. Ainsi, les termes d'indexation sont sélectionnés parmi les termes du corpus qui sont en corrélation avec les termes de la requête en plus des termes sélectionnés statiquement comme dans le modèle vectoriel standard.

Par exemple, pour un utilisateur formulant la requête "java", l'ensemble des termes d'indexation sera constitué de termes sélectionnés de manière dynamique (c-à-d les termes qui sont en corrélation avec le mot "java" dans le corpus documentaire, exemple : "langage", "programmation", "objet", "applet", "swing",..), et des termes sélectionnés au préalable de manière statique comme dans le modèle vectoriel standard . Ensuite, contrairement au modèle vectoriel standard, on associe au vecteur requête, des poids aux termes sélectionnés dynamiquement.

Plus formellement, $DIIR$ est un modèle de recherche documentaire étendant le modèle vectoriel standard. Il est défini par le 5-uplet $< X, Q, T, s, f >$, où X représente le corpus documentaire, Q est l'ensemble des requêtes, T représente l'ensemble des termes d'indexation, s est la fonction de similarité et f la fonction de construction des termes d'indexation, tel que pour une requête donnée q nous avons $T = f(X, q)$.

L'algorithme ci-dessous décrit le processus de recherche documentaire associé à une nouvelle requête q.

Algorithm 1: Algorithme de Recherche documentaire
Données:
q : requête utilisateur
X : corpus documentaire
début
1. construction du graphe G de fréquence d'occurrence et de co-occurrence
2. construction de l'ensemble de termes T_{static}
3. **pour** *chaque requête utilisateur* q **faire**
   1. construction de l'ensemble des termes d'indexation $T = T_{static} \bigcup T_{dynamic}$
   2. construction du vecteur de pondération V_{weight}
   3. représentation vectorielle de la requête q et de l'ensemble des documents du corpus sur l'ensemble T des termes d'indexation
   4. enrichissement du vecteur requête q et nouvelle pondération des vecteurs de documents du corpus
   5. calcul de similarité entre le vecteur requête enrichi et l'ensemble des vecteurs de documents
   6. proposition à l'utilisateur des documents les plus similaires à la requête

fin

Dans la suite de cette section, nous décrivons les étapes principales de l'algorithme.

3.1. *Construction du graphe G de fréquence d'occurrence et de co-occurrence*

Afin de pouvoir extraire les termes d'indexation qui sont en corrélation avec la requête, un graphe G de fréquence d'occurrence et de co-occurrence modelisant le corpus X est construit.

Plus formellement $G =< V, E >$ est un graphe étiqueté tel que :

1) $V = \{(t_1, f_1)..(t_n, f_n)\}$ représente l'ensemble des sommets de G, où chaque sommet (t_i, f_i) est représenté par un terme t_i et sa fréquence f_i.

2) $E = \{(t_i, t_j, fco(t_i, t_j))/t_i, t_j \in V\}$ est l'ensemble des arêtes de G, où $fco(t_i, t_j)$ représente la fréquence de co-occurrence entre les termes t_i et t_j.

Deux termes t_i, t_j sont en co-occurrence, s'ils apparaissent en même temps dans le même contexte. Un contexte de co-occurrence peut correspondre à une phrase, un paragraphe, ou même l'ensemble du document [BES 99, MOK 04]. Étant donné que nous considérons que les éléments pertinents sont généralement proches dans un document, nous considérons dans notre modèle qu'un contexte correspond à une phrase. La fréquence de co-occurrence correspond au nombre d'occurrence de cette co-occurrence. Ainsi, dans G, $fco(t_i, t_j)$ représente la fréquence de co-occurrence des termes t_i, t_j dans l'ensemble des phrases des documents du corpus.
Au cours de la construction automatique du graphe G, tous les termes du corpus sont pris en considération (sauf les mots vides), le processus d'extraction des termes se fait par :

1) Identification des termes (segmentation lexicale).

2) Élimination des mots vides (déterminants, articles, prépositions, ..).

3) Réduction des termes en leurs racines (Stemmatisation ou lemmatisation).

3.2. *Construction de l'ensemble T des termes d'indexation*

L'ensemble T des termes d'indexation est composé de $T = T_{static} \bigcup T_{dynamic}$.

1) Les termes de l'ensemble T_{static} sont sélectionnés de manière statique et dépendent seulement de l'ensemble X. L'ensemble des termes T_{static} est construit de manière classique et ces termes sont choisis de manière à être les plus discriminants possible. Dans notre contexte, le critère de sélection des termes T_{static} est la fréquence en documents.

2) Les termes de l'ensemble $T_{dynamic}$ sont sélectionnés de manière dynamique et varient en fonction de la requête (Algorithme 2). Ils sont sélectionnés parmi les termes qui sont en corrélation avec les termes de la requête. Ainsi pour chaque nouvelle requête q, un nouvel ensemble $T_{dynamic}$ est construit.

Algorithm 2: construction de l'ensemble $T_{dynamic}$
Données:
q : requête utilisateur
G : graphe de fréquence d'occurrence et de co-occurrence du corpus documentaire,
β : constante représentant le seuil de sélection des termes, avec $0 \leq \beta \leq 1$
Sortie: l'ensemble $T_{dynamic}$ de termes d'indexation
début

1. $T_{dynamic} \leftarrow \emptyset$;
2. **pour** *chaque terme* t_i *de* q **faire**
   - **pour** *chaque terme* t_j *de* G *tel que* $fco(t_i, t_j) > 0$ **faire**
     - **si** $\frac{(fco(t_i,t_j))^2}{f_{t_i} \times f_{t_j}} > \beta$ **alors**
       - $T_{dynamic} = T_{dynamic} \cup \{t_j\}$

fin

3.3. *Construction du vecteur de pondération* $\overrightarrow{V}_{weight}$

Afin d'enrichir le vecteur requête initial et de pondérer les vecteurs documents du corpus X, un vecteur de pondération $\overrightarrow{V}_{weight} = (w_1..w_{|T|})$ est calculé, où à chaque terme t_i de T un poids w_i est associé.

Dans $\overrightarrow{V}_{weight}$ un poids non nul est affecté aux termes qui sont en corrélation avec les termes de la requête (c-à-d $T_{dynamic}$). Comme le montre l'algorithme 3, le poids w_i du terme t_i dépend de 3 facteurs :
1. La corrélation de t_i avec les termes de la requête (i.e. $\frac{(fco(t_i,t_j))^2}{f_{t_i} \times f_{t_j}}$, où t_j est un terme de la requête).
2. La fréquence en documents des termes de la requête auxquels t_i est en corrélation.
3. Le nombre de termes de la requête avec lesquels t_i est en corrélation.

Algorithm 3: Construction du vecteur de pondération $\overrightarrow{V}_{weight} = (w_1..w_{|T|})$
Données: T : ensemble des termes d'indexation,
G : graphe de fréquence d'occurrence et de co-occurrence du corpus documentaire
β : constante représentant le seuil de sélection des termes, avec $0 \leq \beta \leq 1$,
Sortie: $\overrightarrow{V}_{weight} = (w_1..w_{|T|})$: vecteur de pondération
début

1. $\overrightarrow{V}_{weight} = \{w_i = 0 / i = 1..|T|\}$
2. $rep = \{rep_i = 0 / i = 1..|T|\}$
3. **pour** *chaque terme* t_j *de* q **faire**
   - **pour** *chaque terme* t_i *de* $T_{dynamic}$ *telque* $\frac{(fco(t_i,t_j))^2}{f_{t_i} \times f_{t_j}} > \beta$ **faire**
     - $w_i = w_i + df(t_j) \times \frac{(fco(t_i,t_j))^2}{f_{t_i} \times f_{t_j}}$
     - $rep_i = rep_i + 1$;
4. **pour** $(i = 1..|T|)$ **faire**
   - $w_i = w_i \times \exp(rep_i)$

fin

$df(t_j)$ est la fréquence en documents du terme t_j.

rep_i est le nombre de termes de la requête avec lesquels t_i est en corrélation.

3.4. *Représentation vectorielle de la requête*

La requête q est initialement représentée par un vecteur $\overrightarrow{q}$ indexé sur l'ensemble des termes T et pondéré par *TF-IDF*. Ensuite $\overrightarrow{q}$ est enrichi par le vecteur $\overrightarrow{V}_{weight}$. Ainsi nous obtenons un nouveau vecteur $\overrightarrow{q'}$ pour lequel des poids non nuls sont associés aux termes qui sont en corrélation avec les termes de la requête initiale q (c-à-d $T_{dynamic}$).

$$\overrightarrow{q'} = \alpha \times \frac{\overrightarrow{q}}{|\overrightarrow{q}|} + (1-\alpha) \times \frac{\overrightarrow{V}_{weight}}{|\overrightarrow{V}_{weight}|}$$

$\overrightarrow{q}$: vecteur requête initiale,
$|\overrightarrow{q}|$: norme euclidienne du vecteur $\overrightarrow{q}$,
$\overrightarrow{V}_{weight}$: vecteur pondération,
$|\overrightarrow{V}_{weight}|$: norme euclidienne du vecteur de pondération $\overrightarrow{V}_{weight}$,
α : constante comprise entre $0 \leq \alpha \leq 1$, permettant l'hybridation entre la requête initiale normalisée $\frac{\overrightarrow{q}}{|\overrightarrow{q}|}$ et le vecteur de pondération normalisé $\frac{\overrightarrow{V}_{weight}}{|\overrightarrow{V}_{weight}|}$.

3.5. *Représentation vectorielle des documents*

Chaque document d est initialement représenté par un vecteur $\overrightarrow{d}$ indexé sur l'ensemble des termes T et pondéré par *TF-IDF*. Ensuite un nouveau vecteur $\overrightarrow{d'}$ qui est, la pondération de $\overrightarrow{d}$ par le vecteur $\overrightarrow{V}_{weight}$, est calculé.

Dans $\overrightarrow{d'}$ les poids des termes de $T_{dynamic}$ sont ajustés en fonction de leurs importances (c-à-d leurs corrélations avec les termes de la requête) et les autres termes sont mis à 0.

$$\overrightarrow{d'} = (w_i * w'_i) \text{ avec } i = 1 \text{ à } \| T \|, \text{ tel que } w_i \in \overrightarrow{d} \text{ et } w'_i \in \overrightarrow{V}_{weight}$$

Ensuite, un nouveau vecteur document $\overrightarrow{d''}$ est calculé comme suit :

$$\overrightarrow{d''} = \alpha \times \frac{\overrightarrow{d}}{|\overrightarrow{d}|} + (1-\alpha) \times \frac{\overrightarrow{d'}}{|\overrightarrow{d'}|}$$

L'objectif de $\overrightarrow{d''}$ est de prendre aussi en considération le poids des termes initiaux de la requête et des termes n'appartenant pas à $T_{dynamic}$.

$|\overrightarrow{d}|$: norme euclidienne du vecteur $\overrightarrow{d}$,
$|\overrightarrow{d'}|$: norme euclidienne du vecteur document pondéré $\overrightarrow{d'}$,
α : constante comprise entre $0 \leq \alpha \leq 1$, permettant l'hybridation entre le vecteur document initial normalisé $\frac{\overrightarrow{d}}{|\overrightarrow{d}|}$ et le vecteur pondéré $\frac{\overrightarrow{d'}}{|\overrightarrow{d'}|}$.

La recherche documentaire s'effectue par un calcul de similarité entre le vecteur requête enrichi $\overrightarrow{q'}$ et chaque vecteur document $\overrightarrow{d''}$. Ainsi, les documents les plus similaires à la requête sont proposés à l'utilisateur.

4. Expérimentation

Une expérimentation sur des données réelles provenant de la campagne d'évaluation AMARYLLISS a été effectuée pour mesurer les performances du modèle *DIIR*.

4.1. *Méthode*

4.1.1. *Données*

Les données sont composées de deux corpus de référence :

– OFIL : un ensemble de 11016 articles du journal *Le Monde*, avec 26 thèmes de recherche correspondant à 587 documents pertinents ;

– INSIT : un ensemble de 163308 notes bibliographiques, avec 30 thèmes de recherche correspondant à 1407 documents pertinents.

Les thèmes de recherche sont composés d'un domaine général du thème, d'un titre résumant le thème, d'une question, d'informations complémentaires sur les documents qui sont jugés pertinents pour ce thème et d'un ensemble de mots-clés cernant le thème [1].

4.1.2. *Comparaison*

Nous avons choisi de comparer *DIIR* avec le modèle vectoriel standard VS pondéré par *TF-IDF*.

Le choix des termes d'indexation dans VS est basé sur la fréquence des termes en documents (nombre de documents dans lesquels le terme apparaît), ainsi nous avons retenu les paramètres classiquement utilisés [SAL 75], c-à-d les termes dont la fréquence en documents est comprise entre $N/100$ et $N/10$ (N est le nombre de documents dans le corpus) .

1. Dans nos expériences, tous ces champs sont fusionnés pour former une seule requête

4.2. *Résultats*

L'évaluation des systèmes de recherche documentaire s'effectue en général avec les mesures standards de précision (P) et de rappel (R), où :

$$P = \frac{\textit{nombre de documents pertinents retournés}}{\textit{nombre total de documents retournés}}$$

$$R = \frac{\textit{nombre de documents pertinents retournés}}{\textit{nombre total de documents pertinents}}$$

En variant les valeurs de α et β, nous avons constaté que les valeurs optimales sont comprise entre 0.2 et 0.8 pour α et entre 0.01 et 0.00001 pour β. Pour cette étude expérimentale α et β sont respectivement fixés à 0.5 et 0.0001.

	DIIR	VS
P(5)	0.3769	0.2462
P(10)	0.3115	0.2269
P(15)	0.2846	0.2256
P(20)	0.2615	0.2115
P(30)	0.2115	0.1897
P(100)	0.0992	0.1081
P(200)	0.0598	0.0710

Tableau 1. *DIIR, VS : valeurs de précision sur le corpus OFIL*

	DIIR	VS
P(5)	0.2333	0.0267
P(10)	0.1933	0.0233
P(15)	0.1778	0.0222
P(20)	0.1517	0.0233
P(30)	0.1322	0.0222
P(100)	0.0767	0.0173
P(200)	0.0525	0.0132

Tableau 2. *DIIR, VS : valeurs de précision sur le corpus INSIT*

Les tables 1 et 2 montrent la précision des documents retournés par $DIIR$ et VS respectivement pour les corpus OFIL et INSIT. $P(n)$ représente la précision sur les n premiers documents retournés. Nous pouvons remarquer que la précision des documents retournés par $DIIR$ est nettement supérieure que celle des documents retournés par VS.

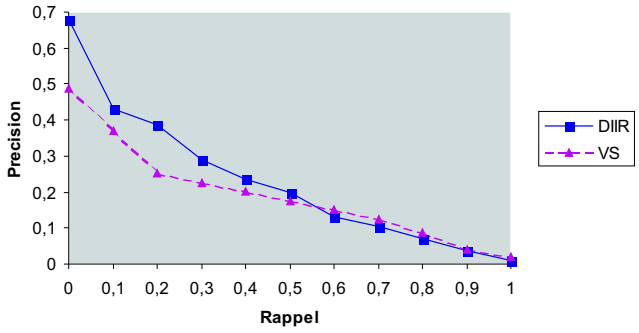

Figure 1. *DIIR, VS : Précision/Rappel pour le corpus OFIL*

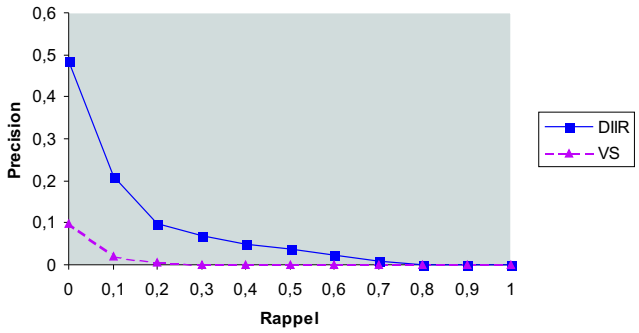

Figure 2. *DIIR, VS : Précision/Rappel pour le corpus INSIT*

Pour mieux illustrer la comparaison entre *DIIR* et *VS*, les Figures 2 et 3 présentent les courbes de *précision/rappel* pour les corpus OFIL et INSIT. Ces résultats montrent que *DIIR* améliore significativement, par rapport à *VS*, le rapport *précision/rappel* pour les faibles valeurs de rappel.

5. Conclusion

Dans cet article, nous avons introduit une nouvelle approche pour la recherche documentaire basée sur une indexation dynamique des documents du corpus. Notre approche permet l'enrichissement de la requête utilisateur et l'adaptation des poids des termes des documents. Cet objectif est atteint par un choix et une pondération dynamique des termes d'indexation. L'étude empirique effectuée sur des données

réelles provenant de la campagne d'évaluation AMARYLLISS confirme la pertinence de notre approche. Une perspective de recherche intéressante serait d'appliquer cette approche à la classification supervisée et non supervisée de documents.

6. Bibliographie

[ATT 77] ATTAR R., FRAENKEL A. S., « Local Feedback in Full-Text Retrieval Systems », *J. ACM*, vol. 24, n° 3, 1977, p. 397–417, ACM Press.

[BES 99] BESANÇON R., RAJMAN M., CHAPPELIER J.-C., « Textual Similarities based on a Distributional Approach », *Proceedings of the Tenth International Workshop on Database And Expert Systems Applications (DEXA99), Firenze, Italy*, 1999, p. 180–184.

[BLA 85] BLAIR D., MARON M., « An Evaluation of Retrieval Effectivness for a Full-Text Document Retrieval System », *Communication of the ACM*, vol. 28, n° 3, 1985, p. 289-299.

[BUC 94] BUCKLEY C., SALTON G., ALLAN J., SINGHAL A., « Automatic Query Expansion Using SMART : TREC 3 », *Text REtrieval Conference*, 1994, p. 0-.

[CRO 97] CROFT W. B., HARPER D. J., *Using probabilistic models of document retrieval without relevance information*, Morgan Kaufmann Publishers Inc., 1997.

[DEE 90] DEERWESTER S., DUMAIS S., LANDAUER T., FURNAS G., HARSMAN R., « Indexing by latent semantic analysis », *Journal of the Society for Information Science, vol. 41*, , 1990, p. 391–407.

[JIN 94] JING Y., CROFT W. B., « An association thesaurus for information retrieval », *Proceedings of RIAO-94, 4th International Conference "Recherche d'Information Assistee par Ordinateur"*, New York, US, 1994, p. 146–160.

[JON 70] JONES K. S., D.M.JACKSON, « The use of automatically-obtained keyword classifications for information retrieval », *Information Processing and Management*, vol. 5, 1970, p. 175–201.

[LAN 98] LANDAUER T., FOLTZ P., LAHAM D., « An Introduction to Latent Semantique Analysis », *Discourse Process*, vol. 25, 1998, p. 259–284.

[LUH 58] LUHN H., « The automatic creation of literature abstracts », *IBM Journal of Research and Development*, vol. 2, 1958, p. 159–165.

[MOK 04] MOKRANE A., AREZKI R., DRAY G., PONCELET P., « Cartographie Automatique du Contenu d'un Corpus de Documents Textuels », *Proceeding of the 7th international conference on the statistical analysis of textual data JADT, 12-15 mars 2004, Louvain-La-Neuve, Belgique*, 2004.

[QIU 93] QIU Y., FREI H.-P., « Concept-based query expansion », *Proceedings of SIGIR-93, 16th ACM International Conference on Research and Development in Information Retrieval*, Pittsburgh, US, 1993, p. 160–169.

[RIJ 79] RIJSBERGEN C. J. V., *Information Retrieval*, Butterworth-Heinemann, 1979.

[ROC 71] ROCCHIO J., « Relevance Feedback in Information Retrieval », *In G. Salton, the SMART Retrieval System : Experiments in Automatic Document Processing*, , 1971, p. 313–323.

[SAL 75] SALTON G., YANG C., YU C., « A Theory of term Importance in Automatic Text Analysis », *Information Processing and Management*, vol. 24, 1975, p. 513–523.

[SAL 83] SALTON G., MCGILL M., « Introduction to Modern Information Retrieval. », *New York : McGraw-Hill*, , 1983.

[XU 00] XU J., CROFT W. B., « Improving the effectiveness of information retrieval with local context analysis », *ACM Trans. Inf. Syst.*, vol. 18, n° 1, 2000, p. 79–112, ACM Press.