Un modèle sémantique pour l'indexation de documents arabes et anglais

Taher ZAKI
Laboratoire IRF-SIC, Université Ibn Zohr Agadir, Maroc
LITIS EA 4108, Université de Rouen, France

Abdellatif ENNAJI
LITIS EA 4108, Université de Rouen, France

Stéphane NICOLAS
LITIS EA 4108, Université de Rouen, France

Driss MAMMASS
Laboratoire IRF-SIC, Université Ibn Zohr Agadir, Maroc

Résumé : Nous présentons ici un système pour l'indexation contextuelle et sémantique de documents en langue arabe et anglais, en se basant sur le voisinage sémantique des termes et l'utilisation d'une modélisation à base radiale. L'usage des graphes et les dictionnaires sémantiques améliore considérablement le processus de l'indexation. Dans ce travail, nous avons proposé une nouvelle mesure TFIDF-Okappi-ABR qui tient en compte la notion de voisinage sémantique à l'aide d'un calcul de similarité entre termes en combinant le calcul du TF-IDF-Okappi avec une fonction noyau à base radiale afin d'identifier les concepts pertinents qui représentent le mieux un document. Des résultats préliminaires et prometteurs sont données sur 2 bases de textes de presse en langue Arabe et Anglaise qui montrent de très bonnes performances par rapport à la littérature.

Mots-clés : Dictionnaire, fonction noyau, formule d'Okappi, graphe sémantique, indexation, TF-IDF, voisinage sémantique.

1. Introduction

La grande masse d'informations textuelles publiées sur le réseau mondial exige la mise en œuvre de techniques efficaces pour l'extraction d'informations pertinentes contenues dans de grand corpus de textes. Le but de l'indexation est de créer une représentation permettant de repérer et retrouver facilement l'information dans un ensemble de documents.

On utilise cette indexation le plus souvent dans les systèmes de recherche d'informations, mais cette indexation peut également servir à comparer et classifier des documents, proposer des mots-clés, faire une synthèse automatique de documents, calculer des co-occurrences de termes… Dans ce papier, nous allons définir un formalisme statistique pour le traitement de documents textuels en arabe et en anglais, et montrer comment ce formalisme peut servir pour le traitement de différentes problématiques telles que l'indexation ou la classification. Notre travail se positionne dans le cadre de la recherche d'information à savoir l'apprentissage statistique qui permet le développement de méthodes génériques utilisables facilement sur différents corpus. Ce formalisme permet d'exploiter à la fois la structure et le contenu textuel de ces corpus.

2. Phase d'indexation

2.1 Problématique

L'indexation est définie comme l'opération qui décrit et caractérise des données résultant de l'analyse du contenu d'un document ou un fragment de document, par des éléments d'un langage documentaire ou naturel en repérant les thèmes présents dans ce document (AFNOR, 1993). L'objectif est de trouver les termes qui caractérisent le mieux le contenu d'un document. Nous nous intéressons donc à la prise en compte des informations explicites autour du texte, à savoir la structure et la répartition des termes, ainsi qu'aux informations implicites, à savoir la sémantique.

2.2 Les étapes du processus

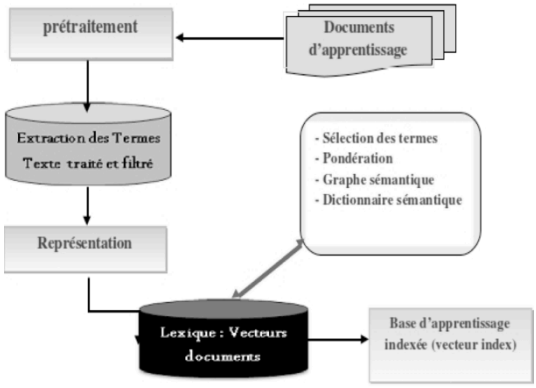


Figure 1 : le Processus de l'indexation

Le système mis au point consiste en 5 étapes fondamentales tel que illustré sur la figure 1 :

A. Base documentaire (Apprentissage et test)
Cette base est un corpus de documents de presse (the Associated Press (AP)) collectée à partir d'internet.

B. Prétraitements (Extraction des termes)
Cette phase consiste à appliquer à l'ensemble du texte une analyse morphologique (lemmatisation, stemming) en premier lieu et un filtrage des termes extraits en deuxième lieu. Ce traitement est nécessaire en raison des variations dans la façon dont le texte peut être représenté en arabe.
La préparation du texte comprend les étapes suivantes :
Les fichiers texte sont converti en codage UTF-16.
Les signes de ponctuation, les signes diacritiques, les non-lettres et les mots vides sont éliminés.
La racinisation des termes restants est opérée à l'aide du stemmer de Khoja (Khoja, 1999) pour les documents arabes, et le stemmer de Porter (porter, 1980) pour les documents anglais.

C. Espace de Représentation
Cette étape permet d'adopter une représentation vectorielle statistique du document à partir des termes retenus pour le représenter. Pour cela, nous avons étendu le modèle vectoriel de Salton en adaptant le calcul du TF-IDF par une combinaison du TFIDF et la formule d'Okappi avec une fonction noyau. Ensuite, pour éviter les problèmes combinatoires liés à la dimension de cet espace de représentation (Sebastiani, 2000) (Deerwester 1990), (Blei, 2003), nous avons adopté une approche de seuillage de fréquence (Document Frequency Thresholding) pour réduire cette dimension.

D. Classification
Pour la phase de classification, nous avons dans cette version préliminaire de notre prototype adopté l'algorithme simple des K plus proches voisins (kppv) pour sa simplicité et pour pouvoir évaluer la pertinence de nos choix de représentation. Nous avons dû également faire le choix d'une métrique adaptée à ce contexte qui est l'operateur de Dice en l'occurrence, dont l'expression est :

$$Dice\ (P_I, P_J) = \frac{2\,|P_I \wedge P_J|}{||P_I| + |P_J||} \quad (1)$$

Où $|P_i|$ est le nombre de termes dans le profile P_i

$|P_i \Lambda P_j|$ est le nombre de termes d'intersection entre les deux profiles P_i et P_j

E. Validations

Pour la validation du prototype, nous avons utilisé une base d'apprentissage très réduite comportant trois thèmes différents (sport, politique, économie et finances). Pour la phase de test, nous avons travaillé sur une base de 400 documents de presse (*Associated Press*) collectés à partir d'internet.

2.3 Pondération des unités index

La manière la plus simple pour calculer le poids d'un terme est de calculer sa fréquence d'apparition car un terme qui apparaît souvent dans un document peut être pertinent pour caractériser son contenu. Plusieurs fonctions de pondération de termes ont été proposées. Nous nous intéressons au classique TF-IDF (term frequency - inverse document frequency) utilisé dans le modèle vectoriel que nous adaptons dans notre travail. Il existe un certain nombre de variantes de TFIDF (Seydoux, 2006). Les critères retenues pour calculer le poids d'un terme sont :

- Une pondération locale qui détermine l'importance d'un terme dans un document. Elle est généralement représentée par sa fréquence (tf).

- Une pondération globale qui détermine la distribution du terme dans la base documentaire. Elle est généralement représentée par l'inverse de la fréquence des documents qui contiennent le terme (idf).

$$a_{ij} = tf(i,j).idf(i) = tf(i,j)\log\left(\frac{N}{N_i}\right) \quad (2)$$

où ***tf(i,j)*** est le *term frequency*, c'est-à-dire le nombre de fois que le terme $\mathbf{t_i}$ apparaît dans le document dj , et idf(i) est l'inverse document frequency, c'est-à-dire le logarithme du rapport entre le nombre N de documents dans le corpus et le nombre Ni de documents qui contiennent le terme ti. Ce schéma d'indexation donne plus de poids aux termes qui apparaissent avec une haute fréquence dans peu de documents.

L'idée sous-jacente est que de tels mots aident à discriminer entre textes de différents thèmes. Le tfidf a deux limites fondamentales : la première est que la dépendance du *term frequency* est trop importante. Si un mot apparaît deux fois dans un document dj, ça ne veut pas nécessairement dire qu'il a deux fois plus d'importance que dans un document d_k où il n'apparaît qu'une seule fois. La deuxième est que les documents plus longs ont typiquement des poids plut forts parce qu'ils contiennent plus de mots, donc les *term frequencies* tendent à être plus élevés. Pour éviter

ces problèmes, nous avons adopté une nouvelle technique d'indexation connue comme la formule d'Okapi (Robertson, 2000) :

$$a_{ij} = \frac{tf\ (i,j)\ .\ idf\ (i)}{[\ (1-b)+b\ \cdot \mathrm{NDL}(\mathrm{d}_j)\ ]+f\ (i,j)} \quad (3)$$

Où $NDL(d_j)$ est la longueur normalisée de $\mathbf{d_j}$, c'est-à-dire sa longueur (le nombre de mots qu'il contient) divisée par la longueur moyenne des documents dans le corpus.

- La mesure N-Gramme

La notion de n-grammes a été introduite par (Shannon, 1948) et est souvent utilisée pour la prédiction d'apparition de certains caractères en fonction d'autres caractères. Les N-Gram sont des séquences de termes dont la longueur est N. Par exemple, l'utilisation des N-Gramm sur le mot « TEXT » est :

bi-grams _T, TE, EX, XT, T_
tri-grams_TE, TEX, EXT, XT_, T__
quad-grams _TEX, TEXT, EXT_, XT__, T___

Les tri-grams pour le mot الممودعيين sont : ا لم, لمو, مود , ودع , دعي, عين

La méthode des N-gramme offre l'avantage d'être une technique indépendante de la langue et permet ainsi une recherche basée sur un segment de mot.

Les systèmes basés sur les n-grammes n'ont pas besoin des prétraitements qui consistent à l'élimination ni des mots vides, ni au Stemming, ni à la lemmatisation, qui sont indispensables pour avoir des performances correctes dans les systèmes à base de recherche de mots (key matching). Pour les systèmes n-grammes, de nombreux travaux ont montré que les performances ne s'améliorent pas en procédant à des traitements d'élimination des "mots vides", de "Stemming" ou de lemmatisation. Nous avons donc mis au point une version à base de N-grammes de notre système pour comparaison.

3. Ressources sémantiques

3.1. Dictionnaire sémantique auxiliaire

Nous avons mis au point un dictionnaire sémantique auxiliaire qui est un dictionnaire hiérarchisé contenant un vocabulaire normalisé sur la base de termes génériques et de termes spécifiques à un domaine. Il ne fournit qu'accessoirement des définitions, les relations entre termes et leur choix l'emportant sur les significations. Les relations communément exprimées dans un tel dictionnaire sont :

les relations taxonomiques (de hiérarchie).

les relations d'équivalence (synonymie).
les relations d'association (relations de proximité sémantique, proche-de, relié-à, etc.).

3.2. Construction du dictionnaire

Le dictionnaire est initialement construit manuellement sur la base des termes retrouvés dans la base d'apprentissage. Mais ce dictionnaire peut être enrichi au fur et à mesure durant la phase d'apprentissage et la classification pour donner plus de flexibilité à notre modèle.
Prenons par exemple le thème finances et économie, le dictionnaire construit est comme suit :

economy, finances, enterprise, industrialism, market, capitalism, socialism, system, brevity, conservation, downsizing, financial status, productive power **…**
finances, budget, account, bill, financing, money, reckoning, score, banking, business, commerce, economic science, economics, political economy, investment **…**
budget, account, bill, calculate, estimate, finance, money, matters, reckon, reckoning, score, assortment, bunch, balanced, cheap,operating budget **…**
….
….

Figure 2 : Dictionnaire de finances et économie

3.3. Les réseaux sémantiques

Les réseaux sémantiques (Quillian, 1968) ont été conçus à l'origine comme un modèle de la mémoire humaine. Un réseau sémantique est un graphe étiqueté (un multigraphe plus précisément). Un arc lie (au moins) un noeud de départ à (au moins) un noeud d'arrivée. Les relations vont des relations de proximité sémantique aux relations partie-de, cause-effet, parent- enfant, etc.
Les concepts sont représentés sous forme de noeuds et les relations sous forme d'arcs. Les liens de différentes natures peuvent être mélangés ainsi que les concepts et instances.
Dans notre système, nous avons utilisé la notion de réseau sémantique comme outils de renforcement du graphe sémantique issu des termes extraits des documents d'apprentissage pour améliorer la qualité et la représentation des connaissances liées à chaque thème de la base documentaire.

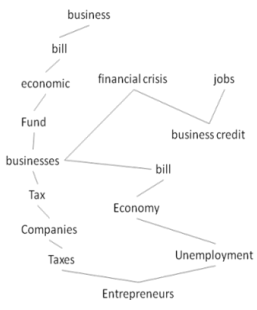
3.4. Construction du graphe

Il est important de noter que l'extraction des termes index se fait dans l'ordre de leur apparition dans le document. Les figures 3 et 4 illustrent ce processus pour un exemple de document du thème finances et économie

WASHINGTON (Reuters) – President Barack Obama signed a $30 billion small **business** lending **bill** into law on Monday, claiming a victory on **economic** policy for his fellow Democrats ahead of November congressional elections. The law sets up a lending **fund** for small **businesses** and includes an additional $12 billion in **tax** breaks for small **companies**."It was critical that we cut **taxes** and make more loans available to **entrepreneurs**," Obama said in remarks at the White House. "So today after a long and tough fight, I am signing a small **business jobs bill** that does exactly that." Obama is trying to show voters, who are unhappy about 9.6 percent **unemployment**, that he and his party are doing everything they can to boost the tepid U.S. **economy**. Democrats said they backed the **bill** because small **businesses** had trouble getting **loans** after the **financial crisis** that began in December 2007. They estimate the **incentives** could provide up to $300 billion in new small **business credit** in the coming years and create 500,000 new **jobs**.	**Business** Bill Economic Fund Businesses Tax Companies Taxes Entrepreneurs Business Jobs Bill Unemployment Economy Bill Businesses loans financial crisis incentives business credit jobs

Figure 3 : texte brute *Figure 4 : Texte après prétraitement et filtrage*

La construction du graphe sémantique tient en compte l'ordre de l'extraction et la distribution des termes dans le document. Chaque terme est associé à une fonction à base radiale qui fixe la proximité à un certain voisinage (zone d'influence sémantique du terme). Ce graphe est ensuite enrichie via le dictionnaire sémantique auxiliaire par l'adjonction de connexions. La correspondance requête- document se fait par une projection des termes de la requête sur le graphe sémantique. Si ces termes sont dans une zone d'influence sémantique forte, alors ce document est pertinent à cette requête. Dans ce qui suit nous allons définir notre fonction à base radiale et nous verrons l'utilité du graphe sémantique pour le calcul de la proximité sémantique entre la requête et le document.



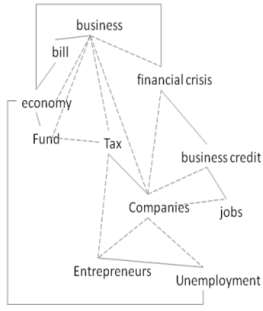


Figure 5 : Graphe Sémantique extrait à partir du document
Figure 6 : Renforcement du Graphe par les connexions sémantiques à partir du dictionnaire auxiliaire

4. Indexation sémantique à fonction à base radiale

Plusieurs travaux ont adapté le modèle vectoriel en indexant directement les concepts à la place des termes. Ces approches traitent essentiellement la synonymie en remplaçant les termes par leurs concepts. Nous traitons des liens plus riches entre les termes en prenant en considération tout les types de relations sémantiques (dans l'idée de construire une ontologie informelle du domaine au sens de conceptualisation). Ceci peut résoudre le problème de la synonymie mais aussi évite les complications causées par les autres relations de spécialisation et de généralisation par exemple.

4.1. Notre contribution pour l'indexation et la classification

Contrairement aux méthodes existantes, nous ne nous restreignons pas à l'utilisation des concepts. En effet, les termes sont enrichis s'ils sont fortement reliés aux concepts voisins et s'ils assurent une bonne connectivité sémantique. Il est important de noter que lors de la recherche, nous pouvons aussi retrouver les termes qui ne sont pas reliés au sein du réseau sémantique.

Pour calculer la similarité entre termes, nous définissons $\phi(d)$ une fonction à base radiale qui associe à chaque terme une zone d'influence caractérisée par le degré de similarité sémantique et la relation entre le terme noyau et ses voisins. (Rada & al., 1989) ont été les premiers à suggérer que la similarité dans un réseau sémantique peut être calculée en se basant sur les liens taxonomiques «is-a». Un moyen des plus évidents pour évaluer la similarité sémantique dans une taxonomie est de calculer la distance entre les noeuds comme le chemin le plus court.

Nous sommes conscients que le calcul de la mesure de similarité par restriction sur le lien «is-a» n'est pas toujours bien adapté car, dans la réalité, les taxonomies ne sont pas toujours au même niveau de granularité, des parties peuvent aussi être plus denses que d'autres. Ces problèmes peuvent être résolus en associant des poids aux liens. Ainsi nous avons choisie de prendre en considération tous les types de relations (problématique conceptuelle) et la répartition des mots dans les documents (problématique structurale).

Nous avons adapté notre système pour qu'il supporte toute sorte de relation sémantique telle que la synonymie, méronymie, hyponimie, taxonimie, antonomie, etc… et nous affectons initialement un poids unité pour les liens sémantiques.

Un réseau sémantique est construit à chaque phase pour modéliser les relations sémantiques entre les termes. Afin d'éviter les problèmes de connexité, nous avons choisi de construire un dictionnaire auxiliaire de telle sorte à avoir une connexité forte du réseau ainsi construit et d'augmenter le poids sémantique des termes descripteurs par la suite.

Dans la section suivante, nous définissons notre mesure TFIDF à base radiale et nous allons voir par la suite comment les poids des termes de l'indexation sont enrichis à partir des sorties de cette mesure.

4.2. Le TF-IDF à base radiale

Les TFIDF à fonction à base radiale (RBF pour Radial Basis Function) s'appuient sur la détermination de supports dans l'espace de représentation E. Cependant, à la différence des TFIDF traditionnels, ceux-ci peuvent correspondre à des formes fictives qui sont une combinaison des valeurs de TFIDF traditionnels, nous les appellerons donc prototypes. Ils sont associés à une zone d'influence définie par une distance (euclidienne, Mahalanobis...) et une fonction à base radiale (Gaussienne, exponentielle...). La fonction discriminante g d'un TFIDF RBF à une sortie est définie à partir de la distance de la forme en entrée à chacun des prototypes et de la combinaison linéaire des fonctions à base radiale correspondantes :

$$g(X) = w_0 + \sum_{i=1}^{N} w_i \, \phi(d(X, \sup_i)) \quad (4)$$

Où $d(x, \sup_i)$ est la distance entre l'entrée x et le support sup_i, $\{w_0, ...,w_N\}$ sont les poids de la combinaison et ϕ la fonction à base radiale.
L'apprentissage de ce type de modèle peut se faire en une ou deux étapes. Dans le premier cas, une méthode de type gradient est utilisée pour ajuster l'ensemble des paramètres en minimisant une fonction objective basée sur un critère comme les moindres carrés. Dans le deuxième cas, une première étape consiste à déterminer les paramètres liés aux fonctions à base radiale (position des prototypes et zones d'influence). Pour déterminer les centres, des méthodes de classification non supervisée sont souvent utilisées. Les poids de la couche de sortie peuvent, dans une seconde étape, être appris par différentes méthodes comme la pseudo-inverse ou une descente de gradient. Dans le cas d'un apprentissage en deux étapes, les TFIDF RBF possèdent alors plusieurs avantages. Par exemple l'apprentissage séparé des fonctions à base radiale et de leur combinaison permet un apprentissage rapide, simple et évite les problèmes de minima locaux (pertinence locale et globale). Les prototypes des TFIDF- RBF représentent la répartition des exemples dans l'espace E de représentation (termes). De plus la gestion des problèmes multi-classes est plus simple dans les TFIDF-RBF. Nous verrons dans la section suivante que les TFIDF RBF sont très semblables sous certaines conditions aux Systèmes d'Inférence Floue.
La modélisation des TFIDF RBF est à la fois discriminante et intrinsèque. En effet la couche de fonctions à base radiale correspond à une description intrinsèque des données d'apprentissage et la couche de

combinaison en sortie cherche ensuite à discriminer les différentes classes.
Dans notre système, nous utilisons des TFIDF RBF avec un apprentissage en deux étapes. La fonction à base radiale est du type fonction de Cauchy de la forme :

$$\phi(d) = \frac{1}{1+d} \quad (5)$$

Et nous définissons deux nouveaux opérateurs :

$$PoidRel(c) = \frac{degré(c)}{nombre\ total\ de\ concepts} \quad (6)$$

C'est le poids relationnel du concept (terme ou vecteur) c et *degré(c)* est le nombre des arrêtes entrantes et sortantes du sommet c. Il représente donc la densité de connexion du concept c au sein du réseau sémantique.

$$DensitéSem(c_1, c_2) = \frac{CoutMin(c_1, c_2)}{Arbre\ recouvrant\ de\ cout\ minimal} \quad (7)$$

DensitéSem(c_1, c_2) est la densité sémantique de la liaison (c_1, c_2). C'est le rapport de la distance sémantique minimale CoutMin(c_1, c_2) entre c_1 et c_2, calculée par l'algorithme de Dijkstra (Cormen et al., 2001). Cette distance est calculée à partir du réseau sémantique ainsi construit à partir de document sur la base du coût minimal de l'arbre recouvrant (c'est l'arbre de coût minimal en suivant tous les chemins minimaux de c_1 vers c_2 et les autres sommets du réseau sémantique). Cette mesure reflète l'importance de la liaison (c_1, c_2) par rapport à l'ensemble des chemins minimaux existants. Par la suite nous calculons la distance sémantique en terme conceptuel comme suit :

$$DistSem(c_1, c_2) = PoidRel(c_1) * PoidRel(c_2) * DensiteSem(c_1, c_2) \quad (8)$$

La mesure de proximité est alors une fonction de Cauchy :

$$Proximité(c_1, c_2) = \frac{1}{1 + DistSem(c_1, c_2)} \quad (9)$$

L'apport de ces opérateurs ainsi définis est qu'ils donnent plus d'importance aux concepts qui ont un voisinage sémantique dense où s'ils ont une bonne connectivité au sein du réseau. Cela a par ailleurs été vérifié durant la validation du prototype.
Nous avons également remarqué que la pondération TFIDF-OKAPPI traditionnelle de quelques termes qui sont considérés comme significatifs pour l'indexation d'un document se trouvent en bas du classement. Après le calcul de la pondération TFIDF-OKAPPI-ABR combinée par

notre fonction à base radiale, ces mêmes termes se retrouvent en haut du classement.
Pour la phase d'indexation, nous allons voir dans la partie qui suit comment les poids des descripteurs index sont générés par la nouvelle mesure à base radiale sur la base de la distance sémantique comme paramètre.

5. Nouvelle pondération des descripteurs index

Les documents sont représentés par des ensembles de vecteurs de termes. Les poids des termes sont calculés en fonction de leur distribution dans les documents. Le poids d'un terme est enrichi par les similarités conceptuelles des termes co-occurrents dans le même thème.
Nous procédons au calcul du TFIDF des termes pour l'ensemble des thèmes de la base d'apprentissage pour en déduire la pertinence globale. On calcule ensuite leur pertinence locale par l'intermédiaire de notre fonction à base radiale définie précédemment en la combinant avec le TFIDF traditionnel et en n'acceptant que les termes situés dans la zone d'influence. Ce poids noté TFIDF-ABR (t) est calculé de la manière suivante :

$$\text{TFIDF-ABR}(t,\text{theme}) = \text{TFIDF}(t,\text{theme}) + \sum_{i=1}^{n} \text{TFIDF}(t_i,\text{theme}) * \varphi(\text{Proximité}(t,t_i)) \quad (10)$$

Avec $\varphi(\text{Proximité}(t, t_i)) < \text{seuil}$

t_i ensemble des n termes dans le thème.

seuil : une valeur qui fixe la proximité à un certain voisinage (zone d'influence sémantique du terme t), nous la fixons dans un premier temps à la proximité entre le concept de t et le **concept contexte** (concept qui représente le thème).

5.1 Okapi à base radiale

Vu les limites de la mesure TFIDF évoquées précédemment, nous avons opté pour un modèle d'Okapi proposé par (Robertson, 2000) en y introduisant une extension sémantique.

Pour ce faire, la fonction $\phi(d)$ calcule le degré de pertinence pour chaque terme au niveau de son voisinage sémantique (zone d'influence). La nouvelle formule devient :

$$a_{ij} = \frac{tf(i,j).idf(i)}{[(1-b) + b \cdot \text{NDL}(d_j)] + f(i,j)} \cdot \phi(d_j) \quad (11)$$

Nous indiquons par $\phi(d_j)$ l'ensemble des termes proches sémantiquement de t_i. Un seuil de similarité est nécessaire pour caractériser l'ensemble de ses éléments. Nous fixons un seuil de similarité pour la valeur de Proximité (t_i,t) qui correspond au degré de similarité

entre t et le concept du thème où il apparaît (le terme est accepté s'il se trouve dans la zone d'influence de terme noyau définie par la fonction à base radiale φ). La relation devient donc :

$$\text{OKAPPI-ABR}(t,\text{theme}) = \text{Okappi}(t,\text{theme}) + \sum_{i=1}^{n} \text{Okappi}(t_i,\text{theme}) * \varphi(\text{Proximité}(t,t_i)) \quad (12)$$

5.2 N-Gramme à base radiale

L'utilisation de la méthode N-gramme (avec N=3 nombre de caractères) dans la recherche des documents arabes est plus efficace que celle du « keyword matching ».
Pour l'indexation et la classification de documents arabes, le choix des mesures statistiques comme les trigrammes et le poids TF*IDF semble pertinent.
L'utilisation de la méthode N-gramme pour l'indexation et la classification des documents reste insuffisante pour obtenir de bons résultats dans la recherche d'information en langue arabe. Pour cela nous avons pensé à ajouter de la pertinence sémantique à cette mesure en tenant compte de la notion du voisinage sémantique des termes extraits par une combinaison N-gramme avec une fonction à base radiale, la formule générale devient :

$$\text{NGRAM-ABR}(t,\text{theme}) = \text{NGRAM}(t,\text{theme}) + \sum_{i=1}^{n} \text{NGRAM}(t_i,\text{theme}) * \varphi(\text{Proximité}(t,t_i)) \quad (13)$$

6. Résultats

Pour la phase d'apprentissage nous avons travaillé sur une base (corpus initial) très réduite de documents étiquetés représentatifs des classes (sport, politique, économie & finance) que l'on cherche à discriminer ou à apprendre et c'est le point fort de notre mesure. Plus cette base est discriminante et représentative plus notre méthode est performante avec de meilleurs résultats.
Pour la phase de test nous avons travaillé sur deux corpus de presse (the Associated Press (AP)) de 400 documents chacun, l'un en langue arabe et l'autre en anglais. Le corpus anglais est une partie extraite d'un corpus plus large de 2246 documents (http://www.cs.princeton.edu/~blei/lda-c/ap.tgz). Pour le corpus arabe, c'est une collection de documents extraite de (www.aljazeera.net).
Le tableau 1 montre les résultats préliminaires obtenus. Ce résultats sont exprimés à travers les critères Rappel, Précision et performances en classification. Ce tableau montre en particulier la pertinence de l'utilisation de notre approche en comparaison avec l'approche N-Grammes.

Corpus	Méthode	Rappel	Précision	Performance en classification (%)
Anglais	TFIDF	0.80	0.83	80.3
	NGRAM	0.56	0.78	65.95
	TFIDF-ABR	0.89	0.92	90.5
	NGRAM-ABR	0.6701	0.8463	74.79
Arabes	TFIDF	0.81	0.81	81.0
	NGRAM	0.45	0.81	57.85
	NGRAM-ABR	0.6341	0.8762	73.57
	Okappi-TFIDF-ABR	0.98	0.98	98.79

Tableau 1 : *Tableau des résultats de l'expérimentation*

7. Conclusion

L'intégration de la notion de voisinage sémantique et de fonctions à base radiale a permis d'améliorer d'une manière très significative les performances de notre système d'indexation indépendamment de la langue manipulée. Ces résultats restent à confirmer sur des corpus plus conséquents, même si de tels corpus sont difficiles à se procurer pour la langue Arabe, qui reste notre objectif primordial.
Nous avons remarqué que les résultats de l'indexation contiennent exactement les mots-clés recherchés triés selon leur pertinence. Nous avons également fixé un seuil pour l'enrichissement sémantique, ce qui peut conduire à retourner quelques termes indésirables assez éloignés de ceux recherchés.
Nous avons aussi constaté que l'hybridation de deux méthodes statistiques améliore considérablement les performances.
Un autre point à prendre en compte et qui peut dégrader la précision des méthodes statistiques traditionnelles, est la présence de concepts complexes. Ce point peut s'avérer une piste intéressante à explorer puisque les concepts longs sont en principe moins sujets à ambiguïté.
Pour réponde à ces différentes situations, nous envisageons l'utilisation d'un algorithme de désambiguïsation et l'hybridation entre différentes mesures en les combinant avec des fonctions noyaux.

Remerciements
Ce travail est soutenu par le Programme Hubert Curien Franco-marocain Volubilis n° MA/10/233 et le projet AIDA du programme Euro méditerranéen 3+3 n ° M/09/05.

Bibliographie

AFNOR (1993). Information et documentation. Principes généraux pour l'indexation des documents. NFZ 47-102.

PORTER M. F. (1980). An algorithm for suffix stripping. *Program*, 14 :130–137, 1980. 15.

SEYDOUX F., RAJMAN M. and CHAPPELIER J.C. (2006**)**. *Exploitation de connaissances sémantiques externes dans les représentations vectorielles en recherche documentaire*. Ph.D. thesis.

BLEI D., NG A., and JORDAN M. (2003). Latent dirichlet allocation.

SEBASTIANI F., SPERDUTI A., and VALDAMBRINI N. (2000). An improved boosting algorithm and its application to automated text categorization. Technical report, Paris, France.

ROBERTSON S., WALKER S., BEAULIEU M.,(2000). Experimentation as a way of life : Okapi at TREC, *InformationProcessing and Management*, vol. 36, no 1,2000,pp. 95-108.

DEERWESTER S., DUMAIS S., FURNAS G., LANDAUER T., and Harshman R (1990). Indexing by latent semantic analysis.

Quillian M.R. (1968). Semantic memory. Semantic information processing, 1968. 65.

RADA R., MILI H., BICKNELL E., BLETTNER M. (1989). « Development and application of a metric on semantic nets », *IEEE Transaction on Systems, Man, and Cybernetics*, vol. 19, no 1, 1989, p. 17–30.

KHOJA S. and GARSIDe S. (1999). Stemming Arabic Text. Computing Department, Lancaster University, Lancaster, U.K. http://www.comp.lancs.ac.uk/computing/users/khoja/stemmer.ps, September 22, 1999.

CORMEN T. H., LEISERSON C. E., RIVEST R. L. and STEIN C. (2001). Introduction à l'algorithmique, (version (en) (ISBN 0-262-03293-7) deuxième édition, 2001, MIT Press and McGraw-Hill, section 24.3, Dijkstra's algorithm, pages 595–601, 2001.

SHANNON, C. (1948). The Mathematical Theory of Communication. Bell System Technical Journal, 27 :379–423 and 623–656.