Pondération des données incertaines dans les systèmes de recherche d'informations : une première approche expérimentale

Caroline Tambellini, Catherine Berrut

Pondération des données incertaines dans les systèmes de recherche d'informations : une première approche expérimentale.

Caroline Tambellini, Catherine Berrut

Laboratoire CLIPS-IMAG

BP 53

38041 Grenoble cedex 9

caroline.tambellini@imag.fr,catherine.berrut@imag.fr

RÉSUMÉ. *Pour trouver l'ensemble des documents répondant à une requête, tout système de recherche d'informations développe une méthodologie formelle ou opérationnelle pour affirmer si oui ou non les termes de chaque document correspondent (plus ou moins partiellement) à ceux de la requête de l'utilisateur. La plupart des systèmes s'appuient sur l'hypothèse que les termes extraits des documents ont été parfaitement reconnus ou identifiés, et de fait leur fonction de correspondance s'appuie sur une capacité à disposer d'une relation d'égalité entre termes du document et termes de la requête. Cependant, il existe des données ou des contextes applicatifs dans lesquels cette hypothèse s'applique difficilement ou bien ne suffit pas pour atteindre un bon niveau qualitatif. Dans ce contexte, nous proposons une première approche où nous modélisons les documents et nous nous posons la question de la pondération des mots.*

ABSTRACT. *To find all documents matching a query, information retrieval systems use a formal or operational method in order to affirm if terms of each document match or not the terms of the user query. The majority of systems are based on the assumption that the terms extracted from the documents were perfectly recognized or identified, so matching is based on equality between document terms and query terms. However, there is some data or context for which this assumption is not easily applicable or is not sufficient to have a good qualitative level. In this context, we propose a first approach in which we model the documents and study the weighting of terms.*

MOTS-CLES : *recherche d'information, données incertaines/ambiguës, fonction de pondération*

KEYWORDS : *information retrieval ,uncertain / ambiguous data, weighting*

1. Introduction / Problématique

Pour trouver l'ensemble des documents répondant à une requête, tout système de recherche d'informations développe une méthodologie formelle ou opérationnelle pour affirmer si oui ou non les termes de chaque document correspondent (plus ou moins partiellement) à ceux de la requête de l'utilisateur. La plupart des systèmes s'appuient sur l'hypothèse que les termes extraits des documents ont été parfaitement reconnus ou identifiés, et de fait leur fonction de correspondance s'appuie sur une capacité à disposer d'une relation d'égalité entre terme du document et terme de la requête. Au niveau de l'indexation des documents, les fondements des systèmes imposent que l'extraction des termes des documents soit sûre : il n'y a pas de doute que tel mot ait été parfaitement reconnu. Ces hypothèses sont par ailleurs celles de la loi de Zipf (Zipf, 1932) et de la conjecture de Luhn (Luhn, 1957) qui sont les fondements des processus d'indexation, et donc des systèmes de recherche d'information.

Les résultats expérimentaux obtenus dans les systèmes de recherche d'informations montrent que cette hypothèse est généralement bonne. Cependant, il existe des données ou des contextes applicatifs dans lesquels ces hypothèses s'appliquent difficilement ou bien ne suffisent pas pour atteindre un bon niveau qualitatif. Prenons deux exemples :

Les informations des documents audio (des conversations par exemple) sont extraites par des outils de reconnaissance de la parole, et ces outils fournissent pour chaque document audio une liste séquentielle de mots liés à la confiance de leur reconnaissance. Traiter ces documents revient donc à travailler avec des documents dont les mots sont incertains et associées à des incertitudes liées au processus d'extraction.

Certains systèmes de recherche d'informations textuelles sont dits orientés précision : leur objectif est de fournir de très bons résultats lors de la mesure de la précision à N documents, avec une valeur N généralement faible (5). Au delà de l'extraction des mots, ces systèmes abordent le problème de l'ambiguïté des mots : savoir si 'porte' est un substantif ou un verbe conjugué dans un document permet d'assurer une meilleure précision face à une requête spécifiant l'une seule des deux possibilités. Les processus linguistiques utilisés dans ces systèmes permettent soit de lever des telles ambiguïtés soit de fournir pour chaque mot la liste de ses catégories morpho-syntaxiques potentielles (Chiaramella, 1986). D'autres processsus fournissent en sortie la catégorie morpho-syntaxique la plus probable pour chaque mot (Treetagger) (Lia_tagg).

Au travers de ces deux exemples, nous voyons qu'il existe effectivement des contextes dans lesquels l'hypothèse qu'un document du corpus est une séquence de mots n'est pas applicable. Dans ces contextes, le document doit être considéré comme une suite de mots associés à des hypothèses d'extraction. De fait, le système

de recherche d'informations qui s'appuie sur de tels documents doit intégrer cette nouvelle dimension.

L'objectif de cet article est de montrer une première approche où nous modélisons de tels documents et les intégrons dans un système de recherche d'informations. Plus spécifiquement, nous nous posons la question de la pondération des mots dans de tels contextes. Les modèles classiques s'appuient sur un décomptage des apparitions des mots dans les documents : ici il faut revisiter cette notion en tenant compte de l'apparition incertaine des mots des documents.

Nous proposons de comparer expérimentalement cette modélisation à des approches classiques. Cette première approche a été faite sur des documents textuels afin de prendre en compte l'ambiguïté des mots. Le corpus d'expérimentations pour les évaluations est issu des données de CLEF-2004. Seule la partie du corpus en français dont les documents sont des articles issus du journal Le Monde est utilisée. Le corpus utilisé est composé de 47 646 documents et de 50 requêtes rédigées en langage naturel.

Après un rappel des fonctions de pondérations classiquement utilisées en recherche d'information, nous présentons notre proposition de fonction de pondération intégrant la dimension d'incertitude. Nous détaillons ensuite la mise en œuvre de cette proposition et les résultats expérimentaux qui en découlent.

2. Pondérations des termes d'un document

2.1. *Notations*

Voici les notations que nous utilisons dans l'article :

N = nombre de documents dans le corpus

tf (t,d) = term frequency = nombre d'occurrences du terme *t* dans le document *d*

totfreq(t) = nombre d'occurrences de *t* dans le corpus = $\sum_{i=1}^{N} tf(t,d_i)$

df (t) = document frequency = le nombre de documents indexés par *t*

2.2. *Fonction générale de pondération*

Une fonction de pondération attribue à chaque terme *t* de chaque document *d* une valeur *w*. Ce poids est calculé en tenant compte de deux grands critères : la force locale *fL* du terme *t* dans *d* et la force globale *fG* de *t* dans le corpus *CO* :

$$w = F\big(fL(t,d), fG(t,CO)\big)$$

La force locale d'un terme dans un document *fL(t,d)* mesure l'importance de ce terme dans le document. La force globale *fG(t,CO)* d'un terme mesure son importance dans le corpus.

Schématiquement, plus un terme est présent dans un document, plus sa force locale *fL* est importante. Plus ce terme est présent dans le corpus, plus sa force globale *fG* est élevée.

fL et *fG* doivent être combinées de manière à évaluer au mieux le poids d'un terme *t* dans un document *d* : la fonction *F* assure la combinaison des deux forces connues pour *t*. *F* doit assurer que plus la force locale d'un terme *t* dans un document *d* est forte, plus *w* doit être élevé, mais aussi que plus la force globale de *t* est élevée, plus *w* doit être faible.

De fait, le poids d'un terme *t* dans un document *d* est généralement calculée par le produit d'une force locale *fL(t,D)* et de l'inverse d'une fonction globale, appelée *ifG(t, CO)* :

$$w = fL(t,d) * ifG(t,CO)$$

Pour définir la force locale et la force globale d'un terme, (Salton, 1971) (Salton and McGill, 1983) (Salton and Buckley, 1988) proposent plusieurs fonctions possibles présentées dans la suite.

2.2.1. *Force locale : importance du terme dans un document*

Comme nous l'avons vu précédemment, la force locale d'un terme est un critère important dans la détermination du poids *w* d'un terme dans un document. Pour représenter cette valeur, le nombre d'occurrences du terme *t* dans le document *d* *tf(t,d)* est communément utilisé.

Nous donnons ici quelques une des fonctions locales les plus couramment utilisées :

- $fL_1(t,d) = tf(t,d)$
- $fL_2(t,d) = \frac{tf(t,d)}{Max(tf(t',d))}$

  où $Max(tf(t',d))$ est la fréquence maximale de l'ensemble des termes de *d*
- $fL_3(t,d) = \log(tf(t,d))$
- $fL_4(t,d) = \log(tf(t,d)+1)$

2.2.2. *Inverse de fonction globale : le pouvoir discriminant du terme*

Nous avons vu précédemment que la force globale d'un terme tient un rôle important dans la définition de la fonction du poids d'un terme : plus un terme a une force globale élevée, plus le poids de ce terme doit être atténué dans les documents.

En effet, un terme souvent présent dans un document et peu présent dans les autres documents sera plus discriminant pour un document qu'un terme apparaissant le même nombre de fois dans le document mais étant présent dans beaucoup d'autres documents. Ceci montre l'intérêt d'utiliser une mesure du pouvoir discriminant (ou non uniformément distribué) du terme dans le corpus de documents.

Cette mesure *ifG* est souvent basée sur *df(t)* (document frequency). D'autres méthodes telles que le ratio signal-bruit ou la valeur de discrimination du terme peuvent être utilisées.

- $ifG_1(t,CO) = \log\left(\frac{N}{df(t)}\right)$
- $ifG_2(t,CO) = discvalue(t)$ = « term discrimination value » mesure via une mesure de similarité entre les documents combien l'utilisation d'un terme augmente (faible discrimination) ou diminue (forte discrimination) la similitude des documents.

$ifG_2(t,CO) = AVGSIM_t - AVGSIM$

Avec $AVGSIM = cte\sum_{i=1,i\neq j}^{N}\sum_{j=1}^{N} similarité(d_i,d_j)$, (par exemple, $cte = \frac{1}{N(1-N)}$)

- $ifG_3(t,CO) = signal(t)$

$signal(t) = \log(totfreq(t)) - noise(t)$ avec $noise(t) = \sum_{i=1}^{N}\frac{tf(t,d_i)}{totfreq(t)}\log\frac{totfreq(t)}{tf(t,d_i)}$

2.3. *Pondérations des termes*

Le poids *w* d'un terme *t* dans un document *d* est calculé comme le produit entre une force locale *fL(t,d)* et une force globale *ifG(t, CO)*. Les formules les plus classiques sont :

- $w_1 = fL_1(t,d) * (ifG_1(t,CO)+1) = tf(t,d) * \log\left(\frac{N}{df(t)}\right)$
- $w_2 = ifL_1(t,d) * ifG_2(t,CO) = tf(t,d) * discvalue(t)$
- $w_3 = ifL_1(t,d) * ifG_3(t,CO) = tf(t,d) * signal(t)$
- $w_4 = ifL_4(t,d) * ifG_1(t,CO) = \log(tf(t,d)+1) * \log\left(\frac{N}{df(t)}\right)$

Les fonctions de pondération que nous venons de présenter ne prennent pas en compte l'incertitude des termes. Se pose alors la question de l'intégration de l'incertitude au sein d'une fonction de pondération.

3. Proposition

3.1. *La donnée incertaine*

Toute donnée extraite par un processus est sujette à être incertaine. Ceci résulte du fait que tout processus peut commettre des erreurs ou ne peut pas décider entre plusieurs possibilités.

Une donnée incertaine apparaît associée à une valeur de certitude : le système est plus ou moins confiant dans sa donnée. Il nous semble intéressant d'exploiter cette valeur de certitude dans les fonctions de pondération.

3.1.1. *La donnée incertaine dans le contexte « étiqueteur syntaxique »*

A chaque terme du document fourni en entrée d'un processus d'étiquetage syntaxique correspond le même terme en sortie du processus associé à une étiquette syntaxique. Pour chaque terme, on a les différentes étiquettes syntaxiques possibles, chacune étant associée à une valeur de certitude (cf. Figure **1**).

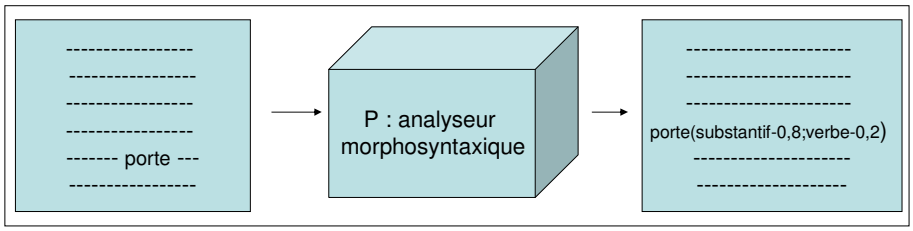

Figure 1. *D'une donnée "simple" à une donnée étiquetée syntaxiquement*

Le processus d'étiquetage syntaxique prend en entrée une donnée « simple » (*terme*) et fournit en sortie une donnée de la forme : *terme (Catégorie$_1$-poidsCatégorie$_1$, Cat$_2$-poidsCat$_2$, ..., Cat$_i$-poidsCat$_i$)*.

Dans notre exemple, en sortie, on a « porte (substantif-0,8 ; verbe-0,2) ». Cette dernière notation signifie qu'il existe une ambiguïté sur le terme porte, la certitude qu'il soit du type substantif commun est de 80% contre 20% pour verbe conjugué.

Dans ce contexte, les mots sont parfaitement identifiées, mais à chaque mot peut être associée une ou plusieurs catégories selon le résultat de l'analyse. Ainsi à la sortie de cette analyse, un document est une liste de mots associés à une ou plusieurs catégories et ce de façon pondérée.

3.1.2. *Notre contexte*

Dans cet article, nous étudions la pondération de ce dernier type : tout document est une séquence de mots, chacun associé à une liste pondérée de catégories. Par exemple, dans la figure 1, le mot 'porte' est associé à la liste ((Substantif, 0,8), (Verbe Conjugué, 0,2)).

Dans un même document, le même mot peut apparaître plusieurs fois. De fait, un mot peut être associé à différentes listes pondérées de catégories. Par exemple, dans un même document, le terme 'porte' peut être trouvé 3 fois, 1 fois en étant associé à la liste (Substantif, 1), une fois à la liste (Verbe Conjugué, 1), et une fois à la liste ((Substantif, 0,8), (Verbe Conjugué, 0,2)).

3.2. *Modélisation des documents*

3.2.1. *Vocabulaire*

Soit $\mathcal{V}$ est un ensemble fermé de Z termes :

$\mathcal{V} = \{t_1, t_2, \ldots, t_Z\}$

3.2.2. *Document*

Un document $\mathcal{D}$ est une séquence de longueur L :

$\mathcal{D} = [t_{d,1}, t_{d,2}, t_{d,3}, \ldots, t_{d,L}]$ avec $t_{d,i} \in \mathcal{V}$.

On peut bien évidemment rencontrer plusieurs fois le même terme dans un document et les $t_{d,i}$ correspondent aux termes dans l'ordre où ils sont rencontrés dans le document.

3.2.3. *Document catégorisé*

3.2.3.1. Catégories

Soit $\mathcal{C}$ un ensemble fermé de C catégories (ex : l'ensemble des catégories morpho-syntaxiques d'un analyseur) :

$\mathcal{C} = \{c_1, c_2, \ldots, c_c\}$

3.2.3.2. Document catégorisé

Un document catégorisé $\mathcal{DC}$ représente le document $\mathcal{D}$ à l'issue de l'étiqueteur syntaxique.

A partir de la représentation du document $\mathcal{D} = [t_{d,1}, t_{d,2}, t_{d,3}, \ldots, t_{d,L}]$ avec $t_{d,i} \in \mathcal{V}$, on définit son document catégorisé $\mathcal{DC}$:

$\mathcal{DC} = [\chi_{d,1}, \chi_{d,2}, \chi_{d,3}, \ldots, \chi_{d,L}]$

Avec $\chi_{d,i} = (t_{d,i}, \{c_j, p_{i,j}\})$ et $t_i \in \mathcal{V}$, $c_j \in \mathcal{C}$, $p_{i,j} \in [0,1]$

$\forall (t, c_i, p_i) \in \chi_i$.

Ce qui signifie que, pour chaque terme de $\mathcal{D}$, on connaît, dans le document $\mathcal{DC}$, l'ensemble complet de toutes ses catégories potentielles.

3.2.4. *Document indexé*

Un document indexé $\mathcal{DI}$ représente le contenu d'un document $\mathcal{D}$, à partir du document catégorisé $\mathcal{DC} = [\chi_{d,1}, \chi_{d,2}, \chi_{d,3}, \ldots, \chi_{d,L}]$

$\mathcal{DI} = \{\chi_i\}$ avec $\chi_i = (t_i, \{c_j, w_{ij}\})$, $t_i \in \mathcal{V}$

Et $\forall \chi_i \in \mathcal{DC}$ – on sait que $\chi_i = (t_i, \{c_j, w_{ij}\})$ - , $\exists$! $\chi_k \in \mathcal{DI}$, $t_i \in \chi_k$

3.3. *Pondération des données incertaines : Calcul des w_{ij}*

Il s'agit ici d'établir le calcul de w_{ij}, le poids du terme t_i avec la catégorie c_j dans le document indexé $\mathcal{DI}$ issu du document $\mathcal{D}$. Pour ce faire, nous travaillons par analogie avec les fonctions locales et globales connues que nous avons présentées précédemment au chapitre 2.2. Dans une première approche, nous prenons comme cadre de travail les fonctions $fL_4(t,d) = \log(tf(t,d)+1)$ et $ifG_1(t,CO) = \log\left(\frac{N}{df(t)}\right)$.

3.3.1. *Notations*

On définit *dc* comme un document de DC : $dc \in \mathcal{DC}$.

- $tf(t,c,dc)$ = le poids des apparitions de *t* avec la catégorie *c* dans le document *dc*

Soit : $tf(t,c,dc) = \sum_{\chi_i \in dc} p_i$ avec $\chi_i \in dc, \chi_i = (t, \{c_j, p_j\}), (t,c,p) \in \chi_i$

- $df(t,c,CO)$ = le nombre de documents contenant *t* avec la catégorie *c*

Soit : $df(t,c,CO) = \|\{D \in CO, \text{ tel que } tf(t,c,DC) > 0\}\|$

3.3.2. *Détermination des valeurs de pondération*

3.3.2.1. Force locale

Par analogie avec la force locale $fL_4(t,d) = \log(tf(t,d)+1)$, on détermine la force locale pour un terme *t* avec une catégorie *c* dans un document *d* comme :

$$fL_5(t,c,d) = \log((tf(t,c,d))+1)$$

3.3.2.2. Force globale

Pour exprimer la force globale d'un terme, nous utilisons la mesure $ifG_1(t,CO)$ adaptée à notre contexte afin de prendre en compte la catégorie du terme :

$$ifG_4(t,c,CO) = \log\left(\frac{N}{df(t,c,CO)}\right)$$

3.3.2.3. Fonction de pondération

Par analogie avec la fonction de pondération w_4, nous proposons d'utiliser une fonction de pondération de type *tf.idf,* que nous noterons w_5 :

$$w_5(t,c,d) = fL_5(t,d) * ifG_4(t,c,CO)$$

4. Expérimentations

4.1. *Collection-test*

4.1.1. *Définition*

La collection-test utilisée pour les évaluations est issue des données de CLEF-2004. Seule la partie de la collection-test en français dont les documents sont des articles issus du journal Le Monde est utilisée. La collection-test utilisée est composée de 47 646 documents et 50 requêtes résolues et rédigées en langage naturel sont disponibles. Ces requêtes sont composées, d'un titre, d'une partie « description », correspondant à une phrase résumant la requête, et d'une partie « narration », correspondant à un court paragraphe détaillant les documents considérés comme pertinents ou non. Voici un exemple de requête :

<top>

<num> C204 </num>

<FR-title> Victimes d'avalanches </FR-title>

<FR-desc> Trouver des informations sur le nombre de morts par avalanche. </FR-desc>

<FR-narr> Les documents pertinents doivent donner des détails sur le nombre de personnes qui meurent à cause des avalanches, que ce soit dans la description de cas spécifiques d'avalanche avec des victimes ou des statistiques générales sur le nombre de morts à cause d'avalanches. </FR-narr>

</top>

Seules les parties « title » et « desc » seront utilisées pour les expérimentations.

4.1.2. *Prétraitements des documents*

Le corpus comporte 47 646 documents qui sont composés en tout de 21 581 650 mots. Le corpus lemmatisé compte 10 979 202 mots. Le vocabulaire final de l'ensemble des documents est composé de 125 981 mots (une fois les mots vides issus d'un anti-dictionnaire enlevés). Ce dernier chiffre est élevé car dans le vocabulaire, les nombres (dates, …) ainsi que notamment les différents noms propres sont conservés.

4.2. *Les systèmes comparés*

Nous comparons notre proposition avec le modèle vectoriel (Salton, 1971). Les requêtes utilisées pour les deux expérimentations sont les mêmes, à savoir les requêtes de CLEF. Pour évaluer notre proposition, ces requêtes sont étiquetées manuellement, ainsi chaque catégorie associée à chacun des termes de la requête est sûre.

La fonction de correspondance de type cosinus est utilisée pour déterminer les documents pertinents selon une requête. Ainsi, la similarité entre le document j et une requête q est définie par :

$$similarité(d_j, q) = \frac{\sum_{i=1}^{n} w_{i,j} w_{i,q}}{\sqrt{\sum_{i=1}^{n} w_{i,j}^2 \sum_{i=1}^{n} w_{i,q}^2}}$$

Pour le modèle vectoriel, la fonction de pondération utilisée est telle que $w_{i,j} = w_4(t,d)$.

Pour notre pondération, afin de prendre en compte le couple (mot, {catégorie, poids de la catégorie}), on a $w_{i,j} = w_5(t,c,d)$.

Pour le traitement des documents dans le cadre de notre proposition, l'étiqueteur syntaxique IOTA (Chiaramella, 1986) est utilisé. Le système IOTA fournit pour chaque mot, l'ensemble de ses catégories possibles. Chaque mot est suivi au minimum d'une catégorie et au maximum de 6 catégories. L'étiqueteur ne fournissant pas de valeurs de certitude associées aux catégories, nous fixons les poids à attribuer à chaque catégorie en fonction du nombre de catégories renvoyées par le système (***Tableau 1***).

Nombre de cat. possibles	Poids cat. en 1ère position	Poids cat. en 2ème position	Poids cat. en 3ème position	Poids cat. en 4ème position	Poids cat. en 5ème position	Poids cat. en 6ème position
1	1					
2	1	0,5				
3	1	0,45	0,45			
4	1	0,4	0,4	0,4		
5	1	0,35	0,35	0,35	0,35	
6	1	0,3	0,3	0,3	0,3	0,3

Tableau 1. *Valeurs de certitude selon la position et le nombre de catégories (p_i)*

4.3. *Evaluation des résultats*

4.3.1. *Modèle vectoriel versus notre proposition*

4.3.1.1. Rappel - précision

Notre fonction de pondération $w_5(t,c,d)$ améliore la précision pour les premiers points de rappel (cf. ***Tableau 2***) par rapport à la fonction de pondération classique $w_4(t,d)$.

	$w_4(t,d)$ ou tf*idf	$w_5(t,c,d)$
0	0,4129	0,4555
0,1	0,3800	0,4066
0,2	0,3710	0,3505
0,3	0,3559	0,3379
0,4	0,3132	0,2892

Tableau 2. *Rappel – précision*

Ce constat correspond à nos attentes : notre système a pour but d'améliorer la précision par l'apport d'informations supplémentaires (les catégories potentielles) pour chaque terme.

4.3.1.2. Mesure ESL

La mesure ESL (Expected Search Length), introduite par (Cooper, 1968) permet d'évaluer le nombre de documents non pertinents devant être lus avant de lire *n* documents pertinents :

$$ESL(n) = j + \frac{i.s}{r+1}$$ avec

n : le nombre de documents pertinents que l'on veut lire

j : le nombre total de documents non pertinents dans les rangs précédents le rang final (c'est-à dire le rang où on se situe lorsque l'on a lu *q* documents pertinents)

r : le nombre de documents pertinents dans le rang final

i : le nombre de documents non pertinents dans le rang final

s : le nombre de documents pertinents requis dans le rang final (en général s=1)

Nous avons appliqué cette mesure au modèle vectoriel ainsi qu'à notre proposition (cf. ***Figure 2***).

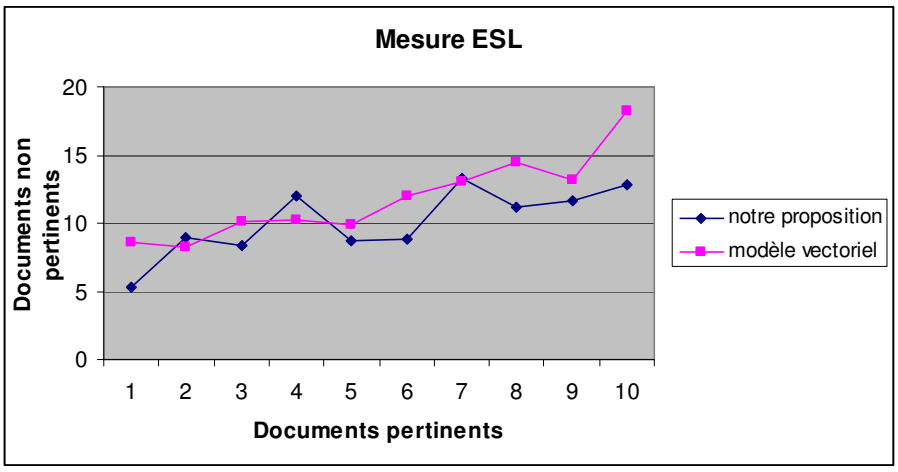

Figure 2. *Mesure ESL : modèle vectoriel vs notre proposition*

Cette mesure fait ressortir le fait que notre proposition améliore les résultats au niveau utilisateur. Ainsi notre proposition améliore non seulement la précision mais également la « facilité » avec laquelle notre système élimine les documents non pertinents.

4.3.2. *Apport mutuel*

Validant expérimentalement que le modèle vectoriel est meilleur au niveau du rappel alors que notre proposition est plus performante au niveau de la précision, nous choisissons d'effectuer une combinaison de ces deux méthodes :

MéthodeCombinée = (1-α) modèle_vectoriel + α notre_proposition

Avec α = 0.7, permettant d'augmenter la précision du système par un poids fort donné à notre proposition.

4.3.2.1. Rappel – Précision

La combinaison des résultats augmente significativement les résultats (cf. ***Tableau 3***). Avec la méthode combinée, on obtient un meilleur rappel.

	$w_4(t,d)$ ou tf*idf	$w_5(t,c,d)$	Méthode combinée
0	0,4129	0,4555	0,4865
0,1	0,3800	0,4066	0,4484
0,2	0,3710	0,3505	0,3924
0,3	0,3559	0,3379	0,3811
0,4	0,3132	0,2892	0,3255
0,5	0,3038	0,2760	0,3136
0,6	0,2696	0,2096	0,2280
0,7	0,2248	0,1346	0,1628
0,8	0,1856	0,1045	0,1292
0,9	0,1324	0,0938	0,1063
1	0,1108	0,0668	0,0768

Tableau 3. *Rappel – Précision*

4.3.2.2. Mesure ESL

Si on reprend la mesure ESL présenté plus tôt, on constate que les résultats sont un peu moins satisfaisants pour la méthode combinée que pour notre proposition (cf. ***Figure 3***). De ce fait, si le but du système de recherche d'information est une précision élevée sur les premiers points de rappel ainsi qu'une bonne élimination des

documents non pertinents, l'utilisation de notre proposition seule donne les meilleurs résultats. Si le but visé est une amélioration des résultats au niveau rappel et précision, la méthode combinée est plus appropriée.

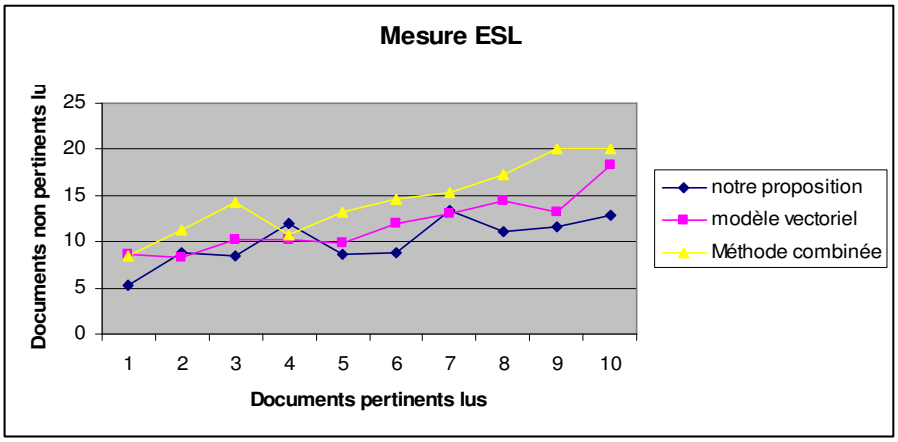

Figure 3.*Mesure ESL*

5. Conclusion et perspectives

Les fondements des processus d'indexation et donc des systèmes de recherche d'informations imposent que l'extraction des termes des documents soit sûre : il n'y a pas de doute que tel mot ait été parfaitement reconnu. Les résultats expérimentaux des systèmes de recherche d'information ont montré que cette hypothèse est généralement bonne. Toutefois, nous avons montré que dans certains contextes, cette hypothèse n'est pas vérifiée : documents audio, documents issus d'étiqueteur syntaxique. Dans de tels contextes, les mots des documents sont associés à des hypothèses d'extraction.

Dans cet article, nous avons montré une première approche de modélisation de ces documents ainsi que leur intégration dans un système de recherche d'informations. Pour cela, nous nous sommes posé la question de la pondération des mots. Les modèles classiques s'appuient sur un décomptage des apparitions des mots dans les documents. Nous avons ici revisité cette notion en tenant compte de l'apparition des mots ainsi que de l'incertitude qui leur est associée.

Après un rapide état de l'art des fonctions de pondérations les plus couramment utilisées dans le domaine de la recherche d'informations, nous avons proposé une pondération adaptée au contexte des données incertaines. Nous avons montré qu'avec peu de variations, il est possible d'adapter ces fonctions afin de prendre en compte l'incertitude des termes et nous avons montré leur apport au sein de

systèmes de recherche d'informations orientés précision. Pour cela, nous avons confronté notre proposition au classique modèle vectoriel.

Nous avons pu mettre en évidence que notre fonction de pondération améliore non seulement la précision sur les premiers points de rappel mais également la « facilité » avec laquelle notre système élimine les documents non pertinents (mesure ESL). Enfin, nous avons montré que la combinaison du modèle vectoriel et de notre proposition permet une amélioration générale du rapport rappel / précision.

Dans des travaux futurs, il nous semble intéressant d'évaluer cette fonction de pondération dans d'autres contextes encore « plus » incertains tel que le domaine de la reconnaissance de la parole. Les bons résultats de notre proposition (basée sur l'intégration de l'incertitude dans les fonctions classiques de pondération) nous encouragent à étudier la mise en place d'une nouvelle fonction de pondération, plus éloignée des fonctions des forces locales et globales du terme classiquement utilisées.

6. Bibliographie

Chiaramella Y. and Defude B. and Bruandet M.F. and Kerkouba D., *IOTA: a full test information retrieval system*, in ACM conference on research and development in information retrieval., Pisa, Italy, pp207-213, september 8, 1986

Cooper W,. *Expected search length: A single measure of retrieval effectiveness based on the weak ordering action of retrieval systems*. American Documentation, 19(1), 30-41, 1968

Lia_tagg, www.lia.univ-avignon.fr

Luhn H., *A statistical approach to mechanized encoding and searching of literary information*, in IBM Journal of Research and Development, 1(4) :309–317, 1957.

Treetagger, www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger

Zipf G., *Selective Studies and the Principle of Relative Frequency in Language*, in Harvard University Press, 1932.