Un modèle de réseau possibiliste pour la recherche d'information

Asma Brini, Mohand Boughanem, Didier Dubois

IRIT 118, Route de Narbonne
31062 Toulouse Cedex 9
{brini,bougha,dubois}@irit.fr

Résumé

Nous proposons un modèle de recherche d'information basé sur les réseaux possibilistes. Les relations de dépendance entre documents-termes d'indexation et termes d'indexation-requête sont quantifiées par des mesures de possibilité conditionnelles. La pertinence d'un document étant donnée la requête est mesurée par deux degrés : la nécessité et la possibilité. La possibilité de pertinence permet d'éliminer dans la liste des documents restitués en réponse à une requête ceux qui ne sont pas pertinents alors que la pertinence nécessaire permet de focaliser sur les documents pertinents. Les expérimentations effectuées sur la collection de tests Le Monde 1994, une sous-collection de $CLEF$ ont permis de montrer l'efficacité de cette approche.

This paper describes an Information Retrieval (IR) model based on possibilistic directed networks. Relations documents-terms and query-terms are modeled through conditional possibility measures rather than a probability measure. The relevance score of a document w.r.t a query is measured by two degrees : the necessity and the possibility. The possibility degree is convenient to filter documents out from the response (retrieved documents) and the necessity degree is useful for document relevance confirmation. Separating these notions may account for the imprecision pervading the retrieval process. Experiments carried out on a sub-collection of $CLEF$, namely Le Monde 1994*, a French newspaper collection, showed the effectiveness of the model.*

Mots-clés *: Recherche d'information, Réseaux possibilistes, Réseaux Bayésiens, Pertinence, Entropie.*

Key-words*: Information Retrieval, Possibilistic Networks, Bayesian Networks, Relevance, Entropy*

1 INTRODUCTION

La Recherche d'Information (RI) consiste à sélectionner dans une collection de documents ceux susceptibles d'être pertinents vis à vis d'un besoin

en information d'un utilisateur. Ce besoin en information est généralement représenté par une requête. La pertinence d'un document vis à vis d'une requête est souvent interprétée par la majorité des modèles de RI, vectoriel [28], probabiliste [21] [14] [27], réseaux d'inférence [36] [1] [34], comme un score calculé à partir des poids des termes du document et ceux de la requête. La pondération de ces termes d'indexation est un des éléments clés dans la mesure de de cette pertinence. Les poids des termes sont, d'une manière générale, obtenus par la combinaison de mesures comme la fréquence d'un terme dans le document (tf), l'importance du terme dans la collection (idf) et la longueur du document (l_d). Ces mesures, obtenues à partir d'informations *pauvres*, résultent souvent d'un point de vue fréquentiste basé sur le comptage des nombres d'apparitions et par conséquent elles ne rendent pas complètement compte de la notion de pertinence, qui reste entachée d'imprécision.

Le modèle que nous proposons dans cet article tente de répondre en partie à ces limites. Tout d'abord, nous proposons d'interpréter la pertinence dans un cadre possibiliste. Ce cadre est plus à même de prendre en compte l'ignorance partielle qui peut affecter les informations utilisées dans les différents calculs. Tout d'abord, le modèle sépare les raisons de sélectionner un document pertinent de celles de le rejeter, en utilisant deux mesures : la nécessité et la possibilité. La possibilité de pertinence tente d'éliminer les documents non pertinents. La nécessité de pertinence met l'accent (le "focus") sur les documents qui semblent très pertinents. Afin de permettre cette interprétation de la pertinence, la pondération des termes dans les documents doit être également réinterprétée. Il a été montré dans [3] [17] que tous les termes d'indexation ne se comportent pas de la même manière dans une collection de documents. Harter fait une distinction entre les mots informatifs appelés aussi mots « spécialisés», qui se focalisent sur un type de documents et les mots non informatifs, non spécialisés qui sont distribués de manière normale sur l'ensemble des documents de la collection. Ceci va dans le sens de notre interprétation de la pertinence ; en effet nous pensons que les termes des documents doivent jouer des rôles différents. Dans un document, il existe des termes fréquents importants (informatifs), nécessaires dans la représentation du document, donc nécessaires, pour décider de la pertinence de ce document vis à vis de la requête, et d'autres termes moins informatifs, qui ne sont que possiblement intéressants pour représenter le contenu du document.

La logique possibiliste offre un bon cadre pour représenter ces deux notions. En effet, notre modèle affecte à chaque terme d'indexation deux valeurs qui traduisent respectivement la certitude et la possibilité qu'un terme d'indexation soit "bon". Le dernier avantage (spécificité) de notre modèle réside dans sa prise en compte explicite de l'absence des termes de la requête dans le document lors de l'évaluation de la pertinence de ce document vis à vis de la requête.

Le papier est organisé comme suit. Nous décrivons tout d'abord dans la section 2 quelques modèles clés de la RI et discutons leur manière d'interpréter la pertinence. La section 3 présente de manière brève quelques notions de base de la théorie des possibilités ainsi que les réseaux possibilistes. La section 4 est consacrée à la description de l'architecture du réseau possibiliste utilisé pour le traitement des requêtes. En section 5, on explique le principe de calcul sur lequel repose le modèle possibiliste pour la RI. La façon dont on agrège les informations relatives aux termes de la requête est détaillée en section 6. La section 7 est consacrée à la définition des pondérations possibilistes reliant les termes et les documents. Enfin, nous discutons dans la section 8 les différentes expérimentations effectuées sur la collection de tests *Le Monde 1994*, une sous-collection de $CLEF$ afin d'évaluer l'intérêt de notre modèle comparativement notamment à $OKAPI$ [27].

2 ETAT DE L'ART

Les modèles de RI proposés dans la littérature peuvent être regroupés en trois catégories selon la théorie sous-jacente à la modélisation de la pertinence :

- modèle vectoriel : la pertinence est définie comme la similarité entre la représentation des documents et de la requête [28] ;
- modèle probabiliste : la pertinence est modélisée par une variable binaire et le but du modèle est d'estimer sa probabilité [27] [14] ;
- modèle logique : la pertinence est reliée à la certitude de pouvoir déduire la requête des représentations des documents ou inversement [38].

Le modèle le plus connu dans la première catégorie est le modèle de Salton [28] [31]. Salton [28] a proposé le système $SMART$ (Salton's Magical Automatic Retriever of Text) basé sur ce modèle. Le sens d'un document est donné par les termes qu'il contient. Les documents et la requête sont représentés par des vecteurs de termes pondérés. La pertinence d'un document vis à vis d'une requête est vue comme une similarité entre les deux vecteurs. La similarité vectorielle peut être mesurée en utilisant des coefficients de type cosinus, Jaccard, Dice [28]. Dans ses premières versions, ce modèle se base uniquement sur la présence-absence des termes dans les représentations (comme dans le modèle Booléen), puis plusieurs pondérations [32][30][29] ont été proposées. La pondération qui a donné les meilleurs résultats a été proposée par Singhal et Buckley, utilisant la normalisation par la méthode de pivot [35]. Ce poids utilise la fréquence d'apparition du terme dans le document (tf), l'importance du terme dans la collection (idf) et la longueur des documents dans lesquels les termes apparaissent.
Concernant la seconde catégorie, plusieurs modèles probabilistes ont été proposés dans la littérature [25] [14] [27]. La principale différence entre ces modèles réside dans la manière d'estimer la probabilité de pertinence. Les

modèles probabilistes peuvent être répartis en deux catégories. La première considère la pertinence comme un concept binaire : soit les documents sont pertinents vis à vis d'une requête soit ils sont non-pertinents. L'appartenance à l'une de ces deux classes est inconnue et les modèles probabilistes tentent de l'estimer. La probabilité de pertinence, notée L, étant donné un document D et une requête Q, notée $P(L \mid D, Q)$ est estimée par un modèle de régression polynomiale [15]. Dans ce modèle, la pertinence est reliée à D et Q. La seconde approche est basée sur la génération de modèle de documents ou de requêtes qui tentent d'estimer $P(D, Q \mid L)$ [26] [14] [20]. Les modèles probabilistes les plus connus sont présentés dans [25] [26] [14]. L'intégration du mélange de $2-Poisson$ a permis de considérer les fréquences des termes [27]. Le modèle de langue proposé dans [23] et développé dans [18] est basé sur des estimations probabilistes. En effet, la pertinence d'un document vis à vis d'une requête dépend de la probabilité de générer une requête à partir du modèle de langue du document. Le document est considéré comme un sous-langage pour lequel un modèle de langue est construit. Ce modèle de langue est obtenu essentiellement à partir des termes du document.

Nous ne présentons dans ce papier que les réseaux Bayésiens vus comme des modèles de la dernière catégorie. Nous considérons ces modèles comme un mélange des modèles logiques et probabilistes. Les modèles les plus connus sont le réseau d'inférence [37] [36] [11] et le modèle de croyance [24] [1] [34]. Pour ces modèles, les documents, les termes d'indexation et la requête sont représentés par des variables binaires et la pertinence est vue comme la déduction des documents pertinents étant donnée une requête. Le modèle de réseaux de croyance généralise les modèles Booléen, vectoriel probabiliste et les réseaux d'inférence. D'autres extensions des modèles de réseaux Bayésiens ont été proposées dans le but d'optimiser les calculs de probabilités conditionnelles en intégrant des relations de dépendance entre des paires de termes ou de documents, ou en traitant des documents hétérogènes. Elles sont proposées dans [12] [9] [8] [16] [10].

Quel que soit le modèle, on voit que la pertinence est vue comme un concept binaire. Cependant, certains travaux de la littérature ont montré que ce concept est graduel et dynamique [25] [33] [19] [5]. De plus, pour tous ces modèles, les termes de la requête absents des documents ne sont pas explicitement considérés dans le calcul des scores de pertinence. Nous proposons pour notre part un modèle qui interprète la pertinence dans un cadre possibiliste. L'approche possibiliste possède à cet égard deux atouts dus à l'emploi de deux évaluations au lieu d'une. D'une part, on peut évaluer de façon plus indépendante les raisons de rejeter un document et les raisons de l'accepter., d'autre part, en cas d'ignorance, on n'est pas obligé de fournir de l'information *a priori*, comme l'exige l'approche Bayésienne. Ces deux avantages sont communs à toutes les approches de l'incertain qui admettent l'imprécision (fonctions de croyance, probabilités imprécises). La théorie des possibilité est la plus simple de ces approches.

3 THÉORIE DES POSSIBILITÉS

La théorie des possibilités introduite par Zadeh [41] et développée par Dubois et Prade [13] évalue l'incertitude sur un ensemble totalement ordonné de valeurs, appelé échelle possibiliste, d'une manière qualitative ou quantitative. Dans le cadre numérique les valeurs des possibilités sur l'intervalle $[0, 1]$ traduisent souvent des bornes supérieures de probabilité. Dans le cadre qualitatif, les valeurs de possibilité ne font que définir un classement des valeurs plus ou moins plausible d'une grandeur. La combinaison conjonctive de distributions de possibilité, exprimée à l'aide de normes triangulaires (t-normes)[13] dépend du cadre formel choisi. Les opérateurs "produit", "minimum" peuvent être utilisés pour combiner des distributions de possibilité indépendantes dans les cadres quantitatif et qualitatif respectivement. Nous nous restreignons, pour nos travaux, au cadre quantitatif.

3.1 Distribution de possibilité

La théorie des possibilités [13] est basée sur les distributions de possibilité. Une distribution de possibilité, notée par π, est une application d'un ensemble d'états possibles X vers l'échelle $[0, 1]$ traduisant une connaissance partielle sur le monde. $\pi(x) = 1$ correspond à un état possible, $\pi(x) = 0$ correspond à un état impossible.
Une distribution de possibilité normalisée exprime qu'un des états est totalement possible, ce qui se traduit par la condition :

$$\max_{x \in X} \pi(x) = 1$$

Si $\max_{x \in X} \pi(x) < 1$, ceci indique une contradiction interne dans la représentation, qui est alors partiellement incohérente.

Mesures de nécessité et de possibilité : Dire qu'un événement est non possible n'implique pas seulement que l'événement contraire est possible mais aussi qu'il est certain. Deux mesures duales sont utilisées : la mesure de possibilité, et la mesure de nécessité. La possibilité d'un événement A, notée $\Pi(A)$ est obtenue par la formule $\Pi(A) = \max_{x \in A} \pi(x)$ et reflète la situation la plus normale dans laquelle A est vraie. Soit $\overline{A}$ le complémentaire de A. La nécessité, notée $N(A)$, d'un événement A, définie par la formule $N(A) = \min_{x \notin A}(1 - \pi(x)) = 1 - \Pi(\overline{A})$, reflète la situation la plus normale dans laquelle A est faux. La distance entre $N(A)$ et $\Pi(A)$ évalue le niveau d'ignorance sur A.

3.2 Conditionnement possibiliste

En logique possibiliste, le conditionnement consiste à modifier la distribution de possibilité initiale π à l'arrivée d'une nouvelle information. En fait, on doit restreindre les états possibles à ceux où la nouvelle information est vraie.

Soit C, une sous classe de X, représentant la nouvelle information. La distribution initiale π est remplacée par $\pi' = \pi(. \mid C)$. Dans un cadre quantitatif, les degrés de possibilités des éléments de C sont proportionnellement modifiés. Ainsi,

$$\pi(x \mid_p C) = \frac{\pi(x)}{\Pi(C)} \text{ si } x \in C$$
$$= 0 \text{ sinon} \quad (1)$$

où $\mid_p$ est le conditionnement basé sur le produit. Notons que c'est exactement la même définition qu'en théorie des probabilités : elle préserve la valeur relative des degrés de possibilités des éléments de C. La seule différence est que $\Pi(C)$ est calculée avec la règle du maximum et non la somme.

3.3 Réseaux possibilistes

Les travaux existants sur les réseaux possibilistes sont soit des adaptations directes de l'approche probabiliste [2], ou des méthodes d'apprentissage à partir de données imprécises [4]. Un graphe possibiliste orienté sur un ensemble de variables $V = V_1, V_2, , , V_N$ est caractérisé par une composante qualitative et une composante numérique. La première est un graphe acyclique orienté comme pour les réseaux Bayésiens. La structure du graphe représente l'ensemble des variables ainsi que l'ensemble des relations d'indépendance. La seconde composante quantifie les liens du graphe en utilisant les distributions de possibilité conditionnelles de chaque noeud dans le contexte de ses parents. Ces distributions de possibilité doivent vérifier la contrainte de normalisation. Pour chaque variable V_i :
(i) Si V_i est un noeud racine et dom_{V_i} le domaine de V_i, la possibilité *a priori* de V_i doit satisfaire $\max_{v_i} \Pi(v_i) = 1, \forall v_i \in dom_{V_i}$
(ii) Si V_i n'est pas un noeud racine, la distribution conditionnelle de V_i dans le contexte de ses parents doit satisfaire $\max_{v_i} \Pi(v_i/PAR_{V_i}) = 1$, $\forall v_i \in dom_{V}i$ où dom_{V_i} est le domaine de V_i, et PAR_{V_i} est l'ensemble des parents de V_i.
Un graphe possibiliste basé sur le produit, noté par GPP, est un graphe possibiliste où les possibilités conditionnelles sont obtenues par le conditionnement de type produit. La distribution de possibilité des réseaux possibilistes basés sur le produit, notée par Π_P, est obtenue par la règle de chaînage

$$\Pi_P(V_1, .., V_N) = PROD_{i=1..N} \Pi(V_i/PAR_{V_i}) \quad (2)$$

où $PROD$ est l'opérateur produit.
Nous proposons, dans ce qui suit une nouvelle approche utilisant les réseaux possibilistes pour traiter les problématiques de la RI.

4 ARCHITECTURE DU MODÈLE

L'approche que nous proposons utilise des réseaux possibilistes orientés. D'un point de vue qualitatif, les documents, les termes d'indexation et la requête sont des variables binaires représentées par des noeuds. Les relations de dépendance entre ces noeuds sont traduites par des arcs orientés. D'un point de vue quantitatif les arcs sont évalués par des degrés de possibilité. L'architecture générale de ce modèle est illustrée dans la figure (1). Un document D_j est instancié ou pas, prenant ses valeurs dans le domaine $\{d_j, \overline{d_j}\}$. L'instanciation d'un noeud document, $D_j = d_j$ (resp. $\overline{d_j}$) signifie que le document est pertinent (resp. non). Une requête Q prend ses valeurs dans le domaine $\{q, \overline{q}\}$. Seule l'instanciation « positive » nous intéresse, et nous considérons $Q = q$ uniquement (mais nous gardons la notation Q). Le domaine d'un noeud terme d'indexation T_i, est $\{t_i, \overline{t_i}\}$. $(T_i = t_i)$ signifie que le terme t_i est présent dans le document (ou dans la requête) et est donc *représentatif* du contenu en information du document (ou de la requête) à un certain degré. Un terme *non-représentatif*, noté par $\overline{t_i}$, est un terme non significatif (éventuellement absent) de la représentation du document ou de la requête.

Soit $\mathcal{T}(D_j)$ (resp. $\mathcal{T}(Q)$) l'ensemble des termes d'indexation du docu-

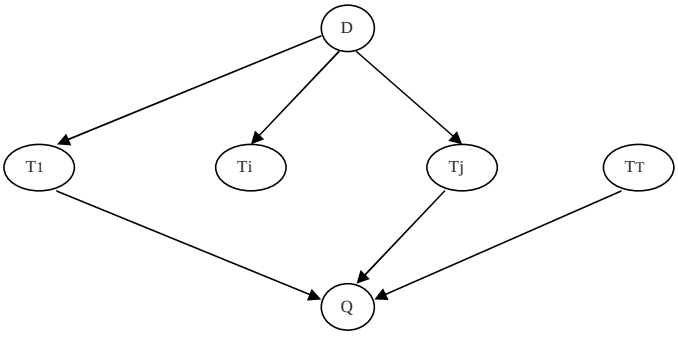

FIG. 1 – Architecture générale

ment D_j (resp. de la requête). La requête exprime la demande de documents contenant certains termes et peut également en exclure d'autres. Les arcs sont orientés des noeuds documents vers les noeuds termes d'indexation définissant les relations de dépendance existantes entre les termes d'indexation et les documents. Les valeurs prises par les termes d'indexation dépendent de l'instanciation des noeuds documents (parents). L'instanciation de la requête propage de l'information uniquement à travers ses termes. Les arcs sont ainsi orientés depuis les noeuds termes d'indexation vers le noeud requête. Les termes apparaissant dans la requête utilisateur forment l'ensemble des parents de Q dans le graphe. Il existe une instanciation de

l'ensemble des parents de la requête ($Par(Q)$) qui représente la requête dans sa forme la plus stricte (conjonctive). Soit θ^Q cette instanciation [1]. Toute instanciation des parents de Q est notée θ. Nous montrerons plus loin dans l'article la façon d'évaluer les arcs. Nous ne considérons pas les relations de dépendance entre couples de termes ici. Cependant ce type de relations pourrait être une information supplémentaire intéressante à exploiter. Ces relations sont exprimables aisément au moyen des réseaux.

5 LE MODÈLE POSSIBILISTE

La principale idée de notre approche concerne l'interprétation de la pertinence. Nous adoptons une approche possibiliste dans le but de mesurer par deux évaluations le score de pertinence d'un document étant donnée une requête. En choisissant l'approche possibiliste, nous cherchons à pouvoir restituer les documents nécessairement ou au moins possiblement pertinents étant donnée une requête. Ce modèle devrait être capable d'inférer des propositions telles que :

- Il est plausible à un certain degré que le document soit pertinent étant donnée la requête, notée par $\Pi(D \mid Q)$;
- Il est certain (dans le sens possibiliste) que le document soit pertinent étant donnée la requête, notée par $N(D \mid Q)$.

Le premier type de proposition est censé éliminer les documents non pertinents. Le second se focalise sur le renforcement de la certitude de la pertinence. Ainsi, le processus de propagation évalue les degrés de possibilité, $\Pi(d_j \mid Q)$, et de nécessité, $N(d_j \mid Q)$. Comme indiqué dans [13],[2] :

$$\Pi(d_j \mid Q) = \frac{\Pi(Q \wedge d_j)}{\Pi(Q)} \tag{3}$$

$$N(d_j \mid Q) = 1 - \Pi(\overline{d_j} \mid Q) = 1 - \frac{\Pi(Q \wedge \overline{d_j})}{\Pi(Q)} \tag{4}$$

La possibilité de Q est donnée par :

$$\Pi(Q) = \max(\Pi(Q \wedge d_j), \Pi(Q \wedge \overline{d_j})). \tag{5}$$

L'équation 3 applique la définition (1) de la possibilité conditionnelle, et l'équation 4 résulte de la dualité entre possibilité et nécessité. L'équation 5 applique la propriété caractéristique de la mesure de possibilité. On en déduit :

$$\Pi(d_j \mid Q) = \min(1, \frac{\Pi(Q \wedge d_j)}{\Pi(Q \wedge \overline{d_j})});\ \Pi(\overline{d_j} \mid Q) = \min(1, \frac{\Pi(Q \wedge \overline{d_j})}{\Pi(Q \wedge d_j)}). \tag{6}$$

[1] Cette configuration représente les termes tels qu'ils sont instanciés dans la requête

Nous cherchons à calculer la forme de $\Pi(Q \wedge D_j)$ en fonction des pondérations sur le graphe. Etant donnée la topologie du graphe et l'équation 2, on trouve :

$$\Pi(Q \wedge D_j) = \max_{\theta^l \in \Theta} \Pi(Q \mid \theta^l) \cdot \prod_{T_i \in \mathcal{T}(Q) \wedge \mathcal{T}(D_j)} \Pi(\theta_i^l \mid D_j) \qquad (7)$$
$$\cdot \Pi(D_j) \cdot \prod_{T_k \in \mathcal{T}(Q) \setminus \mathcal{T}(D_j)} \Pi(\theta_k^l))$$

Avec :
Θ : les configurations possibles de l'ensemble des parents de Q,
θ^l : une configuration possible de θ. θ_i^l : l'instanciation de T_i dans la configuration θ^l ;
Exemple : Soit la requête Q contenant les termes $\{T_1, T_2\}$. Les instanciations possibles des parents de la requête sont dans ce cas : $\Theta = \{\{t_1 \wedge t_2\}, \{t_1 \wedge \overline{t_2}\}, \{\overline{t_1} \wedge t_2\}, \{\overline{t_1} \wedge \overline{t_2}\}\}$; L'instanciation θ_1^1 du terme T_1 dans la première configuration, $\theta^1 = \{t_1 \wedge t_2\}$, est $\theta_1^1 = t_1$.

$\Pi(Q \wedge D_j)$ est calculée pour $D_j \in \{d_j, \overline{d_j}\}$. Nous remarquons que les termes $T_i \in \mathcal{T}(D_j) \setminus \mathcal{T}(Q)$, présents dans le document mais absents de la requête, ne sont pas instanciés lors des calculs. De plus, les termes de la requête qui indexent les documents, $T_i \in \mathcal{T}(Q) \wedge \mathcal{T}(D_j)$, sont évalués dans le contexte de leurs parents par $\Pi(T_i \mid D_j)$, et séparés des termes de la requête absents des documents, pour lesquels une possibilité marginale est calculée, $\Pi(T_k)$.
A l'issue du processus de propagation, chaque document aura donc une valeur de nécessité et de possibilité de pertinence. Les documents répondant à la requête sont classés selon ces deux pertinences. Les documents sont restitués par ordre décroissant de pertinence nécessaire puis de pertinence possible. En effet, ceux classés en premiers sont les documents qui ont une valeur de nécessité supérieure à 0. Les documents possiblement pertinents sont classés après les documents nécessaires ou se retrouvent en haut de la liste lorsque le système ne trouve pas de documents nécessairement pertinents (les documents ayant des degrés de nécessité de pertinence égale à 0).

Illustration

Afin d'illustrer la manière dont la requête est évaluée, considérons le document $D_1 = \{t_2, t_3, t_5, t_6\}$; et la requête $Q = \{t_2, t_3\}$. Deux mesures sont calculées : $\Pi(d_1 \mid Q) = \frac{\Pi(Q \wedge d_1)}{\Pi(Q)}$, et $N(d_1 \mid Q) = 1 - \Pi(\overline{d_1} \mid Q)$ pour les deux instances de D_1, $D_1 = d_1$ et $D_1 = \overline{d_1}$. De plus, $\Pi(Q) = \max(\Pi(Q \wedge d_1), \Pi(Q \wedge \overline{d_1}))$, peut être facilement calculée. Les seuls facteurs à calculer sont : $\Pi(Q \wedge d_1)$ et $\Pi(Q \wedge \overline{d_1})$. $\Pi(Q \wedge d_1)$ est calculé comme

suit (il n'y a pas de termes de la requête hors du document) :

$$\begin{aligned}\Pi(Q \wedge d_1) = \max(&\Pi(Q \mid t_2 t_3) \times \Pi(t_2 \mid d_1) \times \Pi(t_3 \mid d_1)) \times \Pi(d_1),\\ &\Pi(Q \mid t_2 \overline{t_3}) \times \Pi(t_2 \mid d_1) \times \Pi(\overline{t_3} \mid d_1) \times \Pi(d_1),\\ &\Pi(Q \mid \overline{t_2} t_3) \times \Pi(\overline{t_2} \mid d_1) \times \Pi(t_3 \mid d_1) \times \Pi(d_1),\\ &\Pi(Q \mid \overline{t_2 t_3}) \times \Pi(\overline{t_2} \mid d_1) \times \Pi(\overline{t_3} \mid d_1) \times \Pi(d_1))\end{aligned}$$

Pour évaluer les documents étant donnée la requête, nous avons besoin de calculer chacun des facteurs utilisés dans l'expression (7). Nous décrivons dans ce qui suit les différents traitements de la requête en fonction des configurations de ses termes ainsi que des connecteurs utilisés entre eux.

6 AGRÉGATION DES TERMES DE LA REQUÊTE

La possibilité de la requête étant donnés les termes d'indexation, $\Pi(Q \mid \theta)$, dépend de l'interprétation de la requête. Plusieurs interprétations sont possibles. Les termes de la requête peuvent être connectés par une *conjonction*, une *disjonction*, ou par une *somme probabiliste*, ou encore une *somme probabiliste pondérée*. Ces deux dernières agrégations ont déjà été proposées dans les travaux de Turtle [36].
L'idée majeure de l'agrégation de la requête est de mesurer la conformité d'une configuration possible, en l'occurrence celle trouvée dans un document donné, avec la configuration des termes de la requête. Pour ce faire, pour toute configuration, θ^l de Θ, la possibilité conditionnelle $\Pi(Q \mid \theta^l)$ est spécifiée par des fonctions d'agrégation fusionnant les fonctions de ressemblance élémentaires $\Pi(Q \mid \theta_i^l)$. Chaque $\Pi(Q \mid \theta_i^l)$ reflète l'importance de la conformité entre l'instance θ_i^l du terme T_i avec celle de la requête, θ^Q (définie précédemment).
Le stockage de toutes les configurations possibles des termes de la requête est coûteux en espace et le temps de calcul croît de manière exponentielle avec le nombre de termes parents de la requête. En effet, une requête, Q de domaine binaire, composée de 20 termes de domaines binaires aussi, nécessite 2×2^{20} calculs de configurations possibles. Mais, il est à noter que de manière générale les requêtes des utilisateurs ne dépassent pas 3 mots, ceci réduit donc le nombre de configurations. Lorsque l'utilisateur ne fournit aucune information sur les opérateurs d'agrégation de sa requête, l'unique connaissance disponible est l'importance du terme dans la collection. Cette connaissance est disponible pour chaque terme. Nous donnons dans ce qui suit les différentes techniques que nous proposons pour agréger les termes de la requête.

6.1 Agrégation conjonctive

Pour une requête booléenne, *ET*, le processus d'évaluation restitue les documents contenant tous les termes de la requête. Ainsi on pose,

$$\Pi(Q \mid \theta_i^l) = 1 \text{ si } \theta_i^l = \theta_i^Q$$
$$= 0 \text{ sinon}$$

La possibilité de la requête Q étant donnée une configuration possible, θ^l, de Θ de tous ses parents est donnée par :

$$\Pi(Q \mid \theta^l) = 1 \text{ si } \forall\, T_i \in PAR_Q, \quad \theta_i^l = \theta_i^Q \tag{8}$$
$$= 0 \text{ sinon}$$

Dans l'équation 8, il faut que chaque terme T_i parent de la requête Q soit instancié dans θ comme dans la requête. Les documents pertinents pour ce type de requête sont les documents contenant simultanément tous ses termes. Lorsque les termes de la requête concernent un même sujet, des documents plausiblement ou nécessairement pertinents peuvent être restitués. Cependant, plus les termes de la requête sont nombreux et plus ils traitent de sujets différents, plus il est difficile de restituer des documents. Généralement, ce type de requête est trop strict.

6.2 Agrégation disjonctive et quantifiée

Pour une requête booléenne, *OU*, le document est plus ou moins pertinent s'il contient au moins un terme d'indexation de la requête. La pertinence finale d'un document augmente avec le nombre de termes de la requête présents. La disjonction pure est manipulée en remplaçant $\forall$ par $\exists$ dans la requête conjonctive (8).

$$\Pi(Q \mid \theta^l) = 1 \text{ si } \exists\, T_i \in PAR_Q \text{ tel que } \theta_i^l = \theta_i^Q \tag{9}$$
$$= 0 \text{ sinon}$$

Dans le cas de la disjonction, le système restitue les documents contenant au moins un terme de la requête. Cette interprétation est trop tolérante pour discriminer entre les documents.
Un moyen terme entre interprétation conjonctive et disjonctive est le suivant. Convenons par exemple qu'une requête est satisfaite par un document si elle contient au moins K termes communs avec le document. Nous considérons une fonction croissante, $f(\frac{K(\theta^l)}{n})$, tel que $K(\theta^l)$ est le nombre de termes de la requête instanciés dans une configuration donnée θ^l de PAR_Q sachant que la requête contient n termes. Nous posons $f(0) = 0$ et $f(1) = 1$. Par exemple,

$$f(i/n) = 1 \text{ si } i \geq \frac{K}{n},$$
$$= 0 \text{ sinon} \tag{10}$$

Pour l'agrégation donnée par l'équation (10) il faut qu'au moins K termes de la requête soient en conformité avec θ pour sélectionner un document conforme à θ.
D'une manière générale, f peut être une fonction non booléenne si on rend le seuil K flexible. f est alors un quantificateur flou [40].
L'approche quantifiée pour calculer la possibilité d'une requête Q étant donnée une configuration θ^l de tous ses parents, est donnée par :

$$\Pi(Q \mid \theta^l) = f(\frac{K(\theta^l)}{n}) \tag{11}$$

6.3 Noisy OR

En général, les possibilités conditionnelles $\Pi(Q \mid \theta_i^l)$ ne sont pas des booléens mais peuvent dépendre d'une évaluation appropriée des termes T_i. Les termes présents dans la configuration donnée conforme à la requête sont pondérés. La combinaison des termes de la requête peut être inspirée du « noisy-Or » proposé par Pearl [22] pour les réseaux probabilistes. Ce qui signifie que $\Pi(Q \mid \theta^l)$ est évaluée en termes de possibilités conditionnelles de la forme :

$$\Pi(Q \mid t_i \wedge_{k \neq i} \overline{t_k}) \tag{12}$$

Nous supposons l'hypothèse du monde fermé ou Closed World Assumption (CWA) : $\Pi(Q \mid t_i) = \Pi(Q \mid t_i \wedge_{k \neq i} \overline{t_k})$, ce qui permet de se rapprocher de la modélisation booléenne. Ces évaluations sont combinées en utilisant une somme probabiliste. Alors :

$$\begin{aligned} \Pi(Q \mid \theta^l) = 0 \text{ si } \quad & \nexists\, T_i \in PAR_Q \text{ tel que } \theta_i^l = \theta_i^Q \\ = & \frac{1 - \prod_{i:\theta_i^l = \theta_i^Q} (1 - \pi(Q \mid t_i))}{1 - \prod_{T_k \in Par_Q} (1 - \pi(Q \mid t_k))} \text{ sinon} \end{aligned} \tag{13}$$

Pour pouvoir discriminer entre les documents, plus ce nombre de termes croît, plus l'importance des termes instanciés avec la même valeur que dans la requête croît et plus la pertinence du document aura tendance à croître. Seuls les termes instanciés positivement de la requête, $T_i = t_i$, apparaissent au numérateur. Le numérateur contient les termes de la configuration, dans le document en l'occurrence, ayant la même instanciation positive que dans la requête. La formule 13 permet de faire croître la pertinence finale d'un document donné. En effet, le score de pertinence d'un document donné croît selon le nombre de termes qu'il contient ayant la même instanciation (positive) que dans la requête.
Nous rappelons qu'un des problèmes majeurs des réseaux Bayésiens est l'explosion combinatoire liée aux calculs des probabilités (ou possibilités dans notre cas) conditionnelles. Lorsque le nombre de parents ainsi que

leurs domaines augmentent, le nombre de calculs des possibilités conditionnelles augmente d'une manière exponentielle. Un avantage majeur de ce type d'agrégation (13) est qu'il permet de résorber le problème de l'explosion combinatoire liée au calcul des possibilités conditionnelles.

La quantification de la présence ou l'absence d'un terme de la requête dans le document peut être nuancée. Un terme fréquent dans toute la collection n'augmente pas forcément la pertinence du document étant donnée la requête. Par contre, un terme spécifique peut apporter une plus-value à cette pertinence. Ainsi, plus un terme présent dans un document est spécifique, plus la pertinence du document en réponse à une requête qui contient ce terme augmente. La spécificité dans la littérature a été mesurée par la fréquence inverse du terme. Ainsi, on peut légitimement poser

$$\Pi(Q \mid t_i \wedge_{k \neq i} \overline{t_k}) = \frac{idf_i}{logN} = nidf_i \tag{14}$$

avec $idf_i = log\frac{N}{n_i}$, n_i étant le nombre de documents contenant le terme t_i et N le nombre de documents de la collection.

6.4 Possibilité *a priori* des documents

En absence d'information, la possibilité *a priori* d'un noeud document est uniforme

$$\Pi(d_j) = \Pi(\overline{d_j}) = 1$$

Notons que cette représentation de l'ignorance est indépendante de la taille du corpus, ce qui contraste avec la représentation probabiliste de la même situation, qui sera nécessairement biaisée. Nous pouvons obtenir des connaissances sur les documents étant donnée l'importance de ses termes, sa longueur etc. Cette connaissance peut être donnée par un utilisateur, le profil utilisateur etc. Si nous sommes intéressés par les documents longs, la possibilité *a priori* d'un document instancié à $D_j = d_j$ devient :

$$\Pi(d_j) = \frac{l_j}{\max_{k=1,..,N} l_k} = nl_{d_j} \tag{15}$$

avec l_j la longueur du document d_j en terme de fréquence ; $l_j = \sum_i tf_{ij}$. Plus le document est court, moins il est pertinent. Dans tous les cas, $\Pi(\overline{d_j}) = 1$, si on ne veut pas favoriser le document de manière exagérée.

7 PONDÉRATION DES TERMES D'INDEXATION

Pour évaluer la pertinence plausible et la pertinence certaine d'un document étant donnée une requête, nous avons besoin d'exprimer et de définir

les autres arcs du réseau. Un arc reliant un noeud terme à un noeud document quantifie à quel point le terme est représentatif de ce document. Une absence d'arc entre un terme et un document traduit l'absence du terme en question dans le document. La représentativité des termes est selon notre approche considérée sous deux angles différents mais complémentaires. Nous estimons que la combinaison des facteurs $tf \times idf$ n'est pas l'unique approche permettant de donner un sens à la représentativité d'un terme du contenu informatif d'un document donné. Ces deux facteurs sont définis sur des échelles différentes. Le premier est en rapport avec les termes du document qu'il indexe. Le second facteur dépend des documents de la collection qu'il indexe. Les fréquences des termes d'un document donné sont intéressantes pour mesurer à quel point un document est exhaustif. La fréquence inverse permet de mesurer à quel point un terme est spécifique de la collection.
Nous voulons attribuer des poids aux termes sans induire de perte d'information. L'idéal serait de traiter séparément ces deux types d'information (spécificité et/ou exhaustivité). Dans la littérature, deux théories sont connues par leur capacité d'interpréter sous deux angles une information ou une hypothèse. Ces deux théories sont la théorie de Dempster-Shafer [**?**] et la théorie des possibilités. Dans [7], nous montrons que, dans notre problème, les résultats de l'une peuvent être retrouvés par l'autre et inversement. Nous explicitons dans ce papier la méthode adoptée dans un cadre possibiliste.
Nous montrons dans ce qui suit les techniques que nous avons adoptées pour quantifier les poids des termes indexant les documents. Nous montrons par la suite les raisonnements suivis pour quantifier les termes racines : les termes présents dans la requête et absents dans les documents.

7.1 Pondération des termes indexant les documents

Nous tentons dans notre approche d'exprimer de manière plus complète, comparée aux modèles actuels, la pondération d'un terme. Une unification possible de la notion de représentativité serait : « la représentativité d'un terme par rapport à un document décrirait à quel point le document traite du sujet concerné par le terme ».
De ce fait, dans notre cadre de travail, la théorie des possibilités, nous disposons de deux degrés pour évaluer l'incertitude des propositions.
Nous basons la nécessaire représentativité et la plausible représentativité d'un terme sur les deux postulats suivants :
Postulat 1 : Un terme est plus ou moins *possiblement* représentatif du document s'il apparaît fréquemment dans ce document ;
Postulat 2 : Un terme est plus ou moins *nécessairement* représentatif du document s'il apparaît fréquemment dans ce document et rarement dans les autres documents de la collection.
D'après le $Postulat\ 1$, $\Pi(t_i/d_j)$ peut être estimée à partir de la fréquence

relative *tf* :

$$\Pi(t_i/d_j) = ntf_{ij} = \frac{tf_{ij}}{\max_{\forall_{t_k \in d_j}}(tf_{kj})} \tag{16}$$

Selon cette évaluation, un terme n'apparaissant pas dans un document est de poids 0, donc non compatible avec le document. Si son poids vaut 1, alors il apparaît avec une fréquence maximale et le terme est possiblement représentatif du document. Ce terme est un bon candidat potentiel pour le représenter [2]. Ici, « représentatif » ne doit pas nécessairement être compris dans le sens général. Il signifie, dans ce contexte, « utilisable pour restituer ce document dans de la collection ». Un terme représentatif dans le sens général, est un terme qui peut ne pas être utile, ni d'une grande aide pour restituer un document. Supposons un document de la collection qui traite de la *logique floue*. Le mot « floue » est très représentatif mais est potentiellement non utile s'il ne caractérise pas le document parmi d'autres documents ayant le même sujet (traitant du même domaine).

Un terme discriminant dans la collection est un terme qui apparaît (souvent) dans peu de documents de la collection. Nous supposons qu'un terme discriminant est un terme qui est nécessairement représentatif d'un document dans son contexte et donc contribue certainement à le sélectionner parmi d'autres documents. Nous définissons un degré de nécessaire pertinence, ϕ_{ij}, d'un terme t_i pour représenter un document d_j comme un poids de la forme :

$$\phi_{ij} = \mu_1\left(\frac{N}{n_i}\right) * \mu_2\left(ntf_{ij}\right) \tag{17}$$

où $*$: opérateur produit ;
μ_1, μ_2 : fonctions de normalisation. Par exemple, μ_1 fonction logarithmique, μ_2 fonction identité. Alors :

$$\phi_{ij} = \frac{log\frac{N}{n_i}}{log(N)} \cdot ntf_{ij} \tag{18}$$

Ce degré de nécessaire pertinence s'interprète comme la nécessité qu'un terme implique un document et donc aide à restituer ce document. En notant $\rightarrow$ l'implication matérielle, on pose donc

$$N(t_i \rightarrow d_j) = \phi_{ij} \tag{19}$$

Puisque, $\Pi(\overline{d_j}) = 1$ a priori,

$$\text{alors } \Pi(t_i \mid \overline{d_j}) = \Pi(t_i \wedge \overline{d_j}) = 1 - N(t_i \rightarrow d_j) = 1 - \phi_{ij} \tag{20}$$

$$\text{et } \Pi(\overline{t_i} \mid \overline{d_j}) = 1 \tag{21}$$

Dans le tableau 1, nous résumons les possibilités conditionnelles des termes d'indexation étant donnée l'instanciation du noeud document parent. Les va-

[2] A ce stade, nous laissons de côté les relations entre termes, telle que la synonymie par exemple.

TAB. 1 – Possibilités conditionnelles $\Pi(T_i \mid D_j)$

	d_j	$\overline{d_j}$
t_i	ntf_{ij}	$1-\phi_{ij}$
$\overline{t_i}$	1	1

leurs unité de la seconde ligne permettent de respecter la condition de normalisation des possibilités conditionnelles. De plus, si $\phi_{ij} = 0$ on est dans l'ignorance totale quant au terme dans le contexte où le document n'est pas considéré. Si $ntf_{ij} = 1$, poser $\Pi(\overline{t_i} \mid d_j) = 1$ permet de rester neutre quant à la certitude de pouvoir retrouver ce terme comme représentatif du document.

7.2 Termes racines

Les termes racines sont les termes qui apparaissent dans la requête mais pas dans le document. Lors du processus de propagation ces termes sont instanciés par la requête et notre modèle tient compte de l'absence de ces termes. Ceci est une spécificité de notre modèle, car la majorité des modèles de RI ne considèrent pas explicitement ces termes lors du calcul de la pertinence document-requête.

Dans notre approche, un terme discriminant absent du document pénalise la pertinence de ce document. Nous avons présenté dans [6] un nouveau facteur discriminant, noté ndf_i, utilisant l'entropie de Shanon. Ce facteur est proportionnel à la densité d'un terme t_i dans les documents de la collection. On pose

$$p_{ij} = \frac{tf_{ij}}{lj}$$
$$\text{avec } l_j = \sum_j tf_{ij}$$

où p_{ij} est en quelque sorte la probabilité de parvenir au document d_j par tirage selon la fréquence d'apparition du terme t_i. On vérifie $\sum_{j \in N} p_{ij} = 1$. Le facteur ndf_i est obtenu par :

$$ndf_i = \frac{-\sum_{d_j} p_{ij} log(p_{ij})}{\max_{t_k \in T} - \sum_{d_j} p_{kj} log(p_{kj})}$$

avec T l'ensemble des termes de la collection.
Notons que ndf_i est plus expressif que idf, lequel ne tient compte que de la présence ou l'absence d'un terme dans un document. L'impact de l'absence d'un terme de la requête du document est mesurée dans notre cas par :

$$\begin{aligned} \forall T_i \notin \mathcal{T}(D_j),\ \Pi(\theta_i) &= 1 \text{ si } \theta_i^Q = \overline{t_i} \\ &= 1 - ndf_i \text{ sinon} \end{aligned} \quad (22)$$

Un terme uniformément distribué dans la collection minimise le facteur ndf_i et inversement, le facteur croît si le terme se concentre dans dans un petit nombre de documents.

8 EXPÉRIMENTATIONS ET RÉSULTATS

L'objectif de ces expérimentations est de mesurer les performances et la viabilité de notre approche. Pour ce faire, nous avons utilisé la collection de tests standard *Le Monde 1994* issue du programme $CLEF$. Elle comporte des articles du journal français *Le Monde*. Cette collection est composée de 44013 documents et de 40 requêtes, le tout formant 154 MB de données. Les requêtes sont en fait construites à partir des topics proposés par CLEF. En voici un exemple :

```
<top>
<num> 43</num>
<title> El Nino et le temps </title>
<desc>
Rechercher des documents expliquant le phénomène
El Nino et ses répercussions sur le temps a l'échelle
planétaire (y compris ses effets sur la température,
la pression atmosphérique, les précipitations, etc.).
</desc>
<narr>
Les documents pertinents doivent contenir
des informations sur les effets du phénomène El Nino.
Les interactions entre les océans et l'atmosphère
terrestre relevant du phénomène El Nino sont
à prendre en considération. El Nino est
particulièrement intéressant  dans le Pacifique sud,
car il influence le climat a l'échelle planétaire.
</narr>
</top>
```

Ces topics comportent trois champs : titre, description et narrative. Nous avons utilisé le champ titre pour construire ces requêtes. A titre indicatif la requête qui soumise à notre système est : *El Nino et le temps*. Cette requête subit ensuite un certain nombre de transformations habituelles (stemming, suppression des mots vides) identiques à celles effectuées sur les documents.

8.1 Protocole d'évaluation

L'évaluation est effectuée selon le protocole $TREC$. Plus précisément, chaque requête est soumise au système de RI avec les paramètres fixés. Le

système renvoie les 1000 premiers documents pour chaque requête. Les valeurs de précision à $P5, P10, .., Pr.Ex, Pr.Moy$ sont calculées. La précision au point 5, $P5$, est le ratio des documents pertinents parmi les 5 premiers documents restitués. $Pr.Ex,\ Pr.Moy$ sont les précisions exactes et moyennes respectivement [39]. Les paramètres dans notre système représentent les informations considérées lors du processus de propagation déclenché par la requête (formule 7).

8.2 Le modèle optimal

Nous décrivons dans cette section les instanciations prises par les paramètres du modèle optimal, qui a permis d'obtenir les meilleures performances. Les paramètres ont été fixés pour ce modèle tels que décrits dans le tableau 2.

TAB. 2 – Possibilités conditionnelles et marginales

$\Pi(T_i \mid D_j)$	d_j	$\overline{d_j}$
t_i	ntf_{ij}	$1-\phi_{ij}$
$\overline{t_i}$	1	1

$Noisy\ Or$	t_i	$\overline{t_i}$
Q	$1-nidf_i$	1

$\Pi(T_i)$	$Terme\,racine$
t_i	ndf_i
$\overline{t_i}$	1

$\Pi(D_j)$	$Longueur\ des\ documents$
d_j	nl_j
$\overline{d_j}$	1

Dans le tableau 2 la « longueur des documents » et le « terme racine » sont les possibilités marginales définies pour les documents ($\Pi(D_j)$) et les termes racines ($\Pi(T_k)$) respectivement. La représentativité d'un terme d'un document est mesurée par la possibilité conditionnelle ($\Pi(T_i \mid D_j)$). Les possibilités conditionnelles ($\Pi(Q \mid T_i)$) des termes de la requête sont agrégés par l'opérateur du $NoisyOr$.

La figure 2 présente les valeurs des points de précision obtenues pour les requêtes évaluées. Les précisions exacte et moyenne de ce modèle optimal sont de 0.3661 et 0.3821 respectivement. Nous remarquons dans la figure 2 que l'écart entre les points de précision $P5$ et $P10$ est assez élevé comparé aux écarts entre les autres points de précision pris deux à deux. Une explication possible est que notre approche, grâce à cette notion de nécessité de pertinence, permet de restituer les meilleurs documents en début de liste. Cette approche permet de faire de la « haute précision ».

8.3 Comparaisons avec $OKAPI$

Un des apports de notre approche consiste à modéliser d'une nouvelle manière la pertinence. Cette double mesure de pertinence est censée aider le système dans sa décision concernant les documents à restituer ainsi que

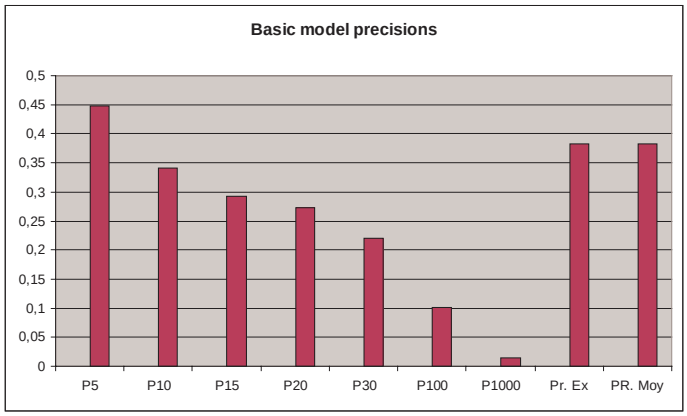

FIG. 2 – Performance du modèle optimal

leur ordre de restitution. Pour ce faire, nous comparons les performances de notre système à un des systèmes les plus performants actuellement à savoir le système $OKAPI$ (BM25) (*BM pour Best Match*) [27][3]. La pertinence d'un document vis à vis d'une requête est calculée dans $OKAPI$ comme suit :

$$RSV(Q,d) = \frac{tf_d(k_1+1)}{k_1 \times ((1-b) + b\frac{l_d}{\Delta l_d}) + tf_d} * log\frac{N-n+0.5}{n+0.5} * \frac{tf_Q \times (k_2+1)}{k_2 \times tf_Q} \quad (23)$$

avec :
$tf_{d(Q)}$: fréquence du terme dans le document (resp. requête), l_d : la longueur du document d ; $l_d = \sum_{i \in L} tf$, les auteurs ont aussi proposé de mesurer en octets les longueurs des documents ; N : nombre de documents dans la collection, n nombre de documents contenant le terme t, Δl_d : la longueur moyenne des documents, $b = 0.75$, $k_2 = 8$, $k_1 = 2$, $k_2 = 8$, $b = 0.75$;

Une première constatation au vu des points de précision est que notre système obtient de meilleures performances. Nous présentons un comparatif des points de précision dans la figure 3.

Nous remarquons une nette amélioration des performances par rapport aux documents restitués en haut de liste. En effet, au vu de ces résultats, il est clair que les valeurs des points de précisions $P5, .., P20$ obtenues par notre système sont plus élevées. Nous obtenons une amélioration de plus de 14% pour la précision à 5 ($P5$). D'une manière générale, comme présenté dans le tableau 3, les précisions $P5, .., P20$ obtenues par l'utilisation de notre

[3]La comparaison a été effectuée en utilisant sur notre index la formule $BM25$.

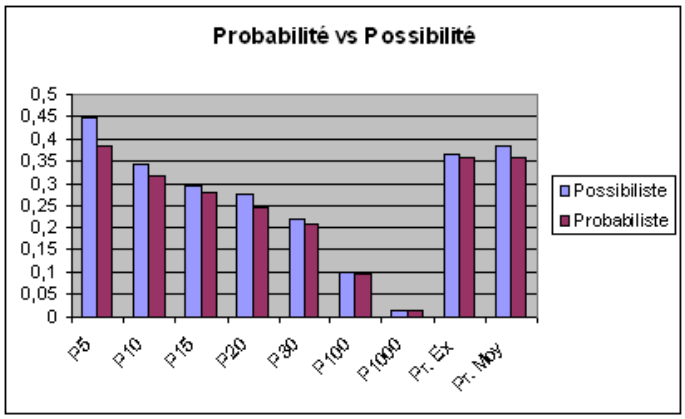

FIG. 3 – Comparatif des deux systèmes : Possibiliste et OKAPI

approche sont supérieures de plus de 5% au modèle $OKAPI$.

TAB. 3 – Pourcentage d'amélioration de notre approche comparée à l'approche probabiliste

	P_5	P_{10}	P_{15}	P_{20}	P_{30}	P_{100}	P_{1000}	$Pr.Moy$
$P_i^{Probabiliste}$	$0,38$	$0,31$	$0,27$	$0,24$	$0,20$	$0,09$	$0,01$	$0,35$
$P_i^{Possibiliste}$	$0,44$	$0,34$	$0,29$	$0,27$	$0,22$	$0,10$	$0,01$	$0,38$
$\%Am$	$16,91$	$7,43$	$4,95$	$11,47$	$6,15$	$4,53$	$2,05$	$8,02$

$P_i^{Possibiliste}$ et $P_i^{Probabiliste}$ désignent la précision au point Pi obtenues respectivement par notre approche et celle d'$OKAPI$. La précision moyenne obtenue par notre système est supérieure de plus de 8% que celle obtenue par $OKAPI$. Nous remarquons aussi que l'augmentation des nombres de documents restitués décroît les précisions de l'approche possibiliste. Parmi les requêtes évaluées par les 2 systèmes, le système possibiliste améliore les précisions à 5 ($P5$) de 14 d'entre elles, et obtient les mêmes valeurs pour 13 d'entre elles. Le système $OKAPI$ obtient de meilleures valeurs $P5$ pour 7 d'entre elles. Intuitivement, notre approche de classement des documents restitués en réponse à une requête utilisateur semble au vu de ces résultats intéressante. Le « découpage » entre les documents certainement (ou nécessairement) pertinents et possiblement pertinents permet de classer les meilleurs documents en haut de la liste.

9 CONCLUSIONS ET PERSPECTIVES

Nous présentons dans ce papier une nouvelle approche de recherche d'information utilisant les réseaux possibilistes. D'une manière générale, la mesure de possibilité permet de filtrer les documents de la liste des documents restitués et la mesure de nécessité permet de donner des raisons de pointer vers un sous-ensemble de documents à restituer. L'originalité de ce travail réside dans le traitement des connaissances disponibles, à savoir la séparation entre les deux notions de représentativité des termes d'indexation (locale dans le contexte du document et globale dans le contexte de la collection) ainsi que la prise en compte de deux variantes de la pertinence. Les expérimentations sur la collection *Le Monde 1994* s'avèrent très encourageants. Les perspectives à court terme concernent l'extension de ce modèle aux documents XML, ainsi que la prise en compte des relations de dépendance existant entre les termes d'indexation et les documents.

RÉFÉRENCES

[1] R.A. Baeza-Yates et B.A. Ribeiro-Neto. *Modern information retrieval*. ACM Press / Addison-Wesley, 1999.

[2] S. Benferhat, D. Dubois, L. Garcia et H. Prade. Possibilistic logic bases and possibilistic graphs. In *Proc. of the Conference on Uncertainty in Artificial Intelligence*, pages 57–64, 1999.

[3] A. Bookstein et D.R. Swanson. Probabilistic models for automatic indexing. *Journal of the American Society for Information Science (JASIS)*, 25 :312–318, 1974.

[4] C. Borgelt, J. Gebhardt et R. Kruse. Possibilistic graphical models. *Computational Intelligence in Data Mining, Courses and Lectures 408, Springer, Wien*, 26 :51–68, 2000.

[5] A.H. Brini et M. Boughanem. Relevance feedback : introduction of partial assessments for query expansion. In *Proc. of the Conference of the european Society for Fuzzy Logic and Technology, (EUSFLAT)*, pages 67–72, 2003.

[6] A.H. Brini, M. Boughanem et D. Dubois. A model for information retrieval based on possibilistic networks. In *Proc. of the symposium on String Processing and Information REtrieval (SPIRE 2005), LNCS, Springer*, pages 271–282, 2005.

[7] Asma H. Brini. *un modèle de recherche d'information basé sur les réseaux possibilsites*. Thèse de doctorat, Université de Toulouse III, Université Paul Sabatier (UPS), 2005.

[8] P.D. Bruza et L.C. van der Gaag. Index expression belief networks for information disclosure. *International Journal of Expert Systems*, 7(2) :107–138, 1994.

[9] S.B. Cousins, J.C. Silverstein et M.E. Frisse. Query networks for medical information retrieval–assigning probabilistic relationships. In *Proc. of the Symposium on Computer Applications in Medical Care (UWSIG)*, pages 803–807. IEEE Computer Society Press, 1991.

[10] F. Crestani, L.M. de Campos, J.M. Fernández-Luna et J.F. Huete. A multi-layered bayesian network model for structured document retrieval. In *Proc. of the 7th European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty (ECSQARU)*, pages 74–86, 2003.

[11] W. Croft et H. R. Turtle. Text retrieval and inference. *Text-Based Intelligent Systems. Current Research and Practice in Information Extraction and Retrieval*, pages 127–155, 1992.

[12] L.M. de Campos, J.M. Fernández-Luna et J.F. Huete. Query expansion in information retrieval systems using a bayesian network-based thesaurus. In *Proc. of the Uncertainty in Artificial Intelligence Conference (UAI)*, pages 53–60, 1998.

[13] D. Dubois et H. Prade. *Possibility Theory*. Plenum, 1988.

[14] N. Fuhr. Probabilistic models in information retrieval. *The Computer Journal*, 35(3) :243–255, 1992.

[15] N. Fuhr. Language models and uncertain inference in information retrieval, 2001. In Proc. of the Language Modeling and IR workshop.

[16] R. M. Fung et B. Del Favero. Applying bayesian networks to information retrieval. *Communications of the ACM (CACM)*, 38(3) :42–48, 1995.

[17] S.P. Harter. A probabilistic approach to automatic keyword indexing. part ii. an algorithm for probabilistic indexing. *Journal of the American Society for Information Science (JASIS)*, 35(3) :280–289, 1975.

[18] D. Hiemstra et W. Kraaj. Twenty-one at trec-7 : Ad hoc and cross language track. In *Proc. of the Text REtrieval Conference (TREC-7)*, pages 227–238, 1998.

[19] J. Kekäläinen et K. Järvelin. Evaluating information retrieval systems under the challenges of interaction and multidimensional dynamic relevance. In *Bruce, H., Fidel, R., P. Ingwersen, P. Vakkari, eds. Emerging Frameworks and Methods, Seattle, Colerado : Libraries Unlimited*, pages 253–270, 2002.

[20] J. Lafferty et C. Zhai. *Probabilistic relevance models based on document and query generation.*, volume 13. Kluwer Academic, 2003.

[21] M. Maron et J. Kuhns. On relevance, probabilistic indexing and information retrieval. *Journal of the Association for Computing Machinery*, 7 :pages 216–244, 1960.

[22] J. Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann San Mateo, Ca, 1988.

[23] J. M. Ponte et W. B. Croft. A language modeling approach to information retrieval. research and development in information retrieval. In *Proc. of the International ACM-SIGIR Conference*, pages 275–281. Proc. of the International ACM-SIGIR Conference, 1998.

[24] B. Ribeiro-Neto et R. R. Muntz. A belief network model for ir. In *Proc. of the International ACM-SIGIR Conference*, pages 253–260, 1996.

[25] C.J. Van Rijsbergen. *Information Retrieval*. Butterworth-Heinemann, Newton, MA, 1979.

[26] S.E. Robertson, C.J. van Rijsbergen et M.F. Porter. Probabilistics models of indexing and searching. *Information retrieval research, (Ed. W.R. Oddy et al), London :Butteworths*, pages 36–65, 1981.

[27] S.E. Robertson et S. Walker. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *Proc. of the International ACM-SIGIR Conference*, pages 232–241, 1994.

[28] G. Salton. *The Smart retrieval system-experiments*. Automatic Document Processing, Prentice Hall Inc, 1971.

[29] G. Salton. Syntactic approaches to automatic book indexing. In *Proc. of the annual meeting on Association for Computational Linguistics (ACL)*, pages 204–210. Department of Computer Science, Cornell University, Ithaca, New York, 1988.

[30] G. Salton et C. Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing & Management (IPM)*, 24(5) :513–523, 1988.

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

[32] G. Salton et C.S. Yang. On the specification of term values in automatic indexing. *In Journal of Documentation*, (29) :351–372, 1973.

[33] T. Saracevic. Relevance reconsidered. In *Information science : Integration in perspectives*, pages 201–218. Proc. of the Conference on Conceptions of Library and Information Science, 1996.

[34] I. Silva, B. A. Ribeiro-Neto, P. Calado, E.S. de Moura et N. Ziviani. Link-based and content-based evidential information in a belief network model. In *Proc. of the International ACM-SIGIR Conference*, pages 96–103, 2000.

[35] A. Singhal, C. Buckley et M. Mitra. Pivoted document length normalization. *Proc. of the International ACM-SIGIR Conference*, 32(2) :21–29, 1996.

[36] H. R. Turtle et W. B. Croft. Evaluation of an inference network-based retrieval model. *ACM Transaction on Information Systems*, 9(3) :7187–222, 1991.

[37] H.R. Turtle et W.B. Croft. Inference networks for document retrieval. In *Proc. of the International ACM-SIGIR Conference*, pages 1–24, 1990.

[38] C. J. van Rijsbergen. A non-classical logic for information retrieval. *In Computer Journal*, 29(6) :481–485, 1986.

[39] Ellen M. Voorhees et Donna Harman. Overview of the ninth text retrieval conference (trec-9). In *TREC*, 2000.

[40] R. R. Yager et H. L. Larsen. Retrieving information by fuzzification of queries. *Journal of Intelligent Information Systems*, 2(4) :106–119, 1993.

[41] L. A. Zadeh. Fuzzy sets as a basis for a theory of possibility. *Fuzzy Sets and Systems*, 1 :3–28, 1978.