Réseaux possibilistes pour un modèle de recherche d'information

Asma BRINI — Mohand BOUGHANEM — Didier DUBOIS

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

RÉSUMÉ. *Nous proposons dans ce papier un modèle de recherche d'information utilisant les réseaux possibilistes. Les relations de dépendance entre documents-termes d'indexation et termes d'indexation-requête sont quantifiées par des mesures de nécessité et de possibilité. La pertinence d'un document étant donnée la requête est mesurée par deux degrés : la nécessité et la possibilité. La requête utilisateur déclenche un processus de propagation dans le but de restituer les documents nécessairement et possiblement pertinents. La possibilité de pertinence permet d'éliminer des documents de la liste des documents restitués 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.*

ABSTRACT. *This paper proposes a model for Information Retrieval (IR) based on possibilistic directed networks. Relations documents-terms and query-terms are modeled through possibility and necessity measures rather than a probability measure. The relevance value for the document given the query is measured by two degrees: the necessity and the possibility. More precisely, the user's query triggers a propagation process to retrieve necessarily or at least possibly relevant documents. 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 LeMonde* 1994*, a French newspapers collection, showed the effectiveness of the model.*

MOTS-CLÉS : *Modèle de recherche d'information, Réseaux Possibilistes, Pertinence,Théorie des possibilités*

KEYWORDS: *Information retrieval model, Possibilistic networks, Relevance, Possibility theory*

1. Introduction

Les travaux que nous proposons s'inscrivent dans la définition d'un nouveau modèle de RI permettant notamment une nouvelle modélisation de la pertinence. Un des objectifs de ce travail est de proposer une approche moins restrictive pour la modélisation de la pertinence. En effet, nous donnons deux sens différents mais complémentaires à la notion de pertinence. Nous proposons une pertinence certaine et une pertinence plausible d'un document étant donnée une requête. Les documents sont restitués par ordre décroissant de leur **nécessaire pertinence** et ensuite, ou à défaut, par ordre décroissant de leur **pertinence plausible ou possible**. Ces pertinences dépendent, comme dans les modèles actuels, des poids des termes de la requête et des documents. Tous les termes de la requête sont pris en compte qu'ils soient absents ou présents dans le document (pour lequel nous calculons la pertinence). Nous proposons un modèle basé sur un réseau possibiliste. Les noeuds représentent les documents, termes d'indexation et la requête. Les relations de dépendance traduisant la représentativité d'un terme dans un document ou une requête sont quantifiées par des degrés de possibilité et de nécessité.
Afin de valider nos propositions nous avons effectué des expérimentations sur une collection de tests standard de RI. Cette collection provient de la campagne d'évaluation $CLEF$ (Cross Language Evaluation Forum). Nous avons évalué l'impact d'une double mesure de pertinence et d'une double mesure de représentativité sur les performances de notre système. Nous avons pour cela comparé notre modèle à un des modèles les plus performants, en termes de rappel et précision, à savoir le modèle probabiliste $OKAPI$. Les résultats obtenus par notre modèle sur les collections utilisées sont supérieurs que ceux du système $OKAPI$ sur la majorité des points de précision considérés.
Nous survolons dans la section 2 quelques notions de théorie des possibilités et de réseaux possibilistes. Nous détaillons dans la troisième section ($section3$) le modèle que nous proposons, sa modélisation graphique ainsi que la quantification des arcs qui relient chaque paire de noeuds. Dans la section 4 nous décrivons les expérimentations effectuées et les résultats obtenus.

2. La théorie des possibilités

La théorie des possibilités introduite par Zadeh [ZAD 78] et développée par Dubois et Prade [DUB 88] traite l'incertitude sur l'intervalle $[0, 1]$, appelé échelle possibiliste, d'une manière qualitative ou quantitative. Nous nous restreignons, pour nos travaux, au cadre quantitatif.
Distribution de possibilité La théorie des possibilités 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.
Mesures de nécessité et de possibilité Dire qu'un événement est non possible n'implique pas seulement que l'événement contraire est possible mais qu'il est certain.

Deux mesures duales sont utilisées : la mesure de possibilité $\Pi(A)$, et la mesure de nécessité $N(A)$. La possibilité d'un événement A, notée $\Pi(A)$ est obtenue par $\Pi(A) = max_{x \in A} \pi(x)$ et décrit la situation la plus normale dans laquelle A est vraie. La nécessité $N(A) = min_{x \notin A} 1 - \pi(x) = 1 - \Pi(\overline{A})$ d'un événement 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.

Conditionnement possibiliste En logique possibiliste, le conditionnement consiste à modifier la distribution de possibilité initiale π à l'arrivée d'une nouvelle information i. Soit C, une sous classe de X, $C = [i]$ l'ensemble des modèles de i. La distribution initiale π est remplacée par $\pi' = \pi(\bullet / C)$. Dans un cadre quantitatif, les éléments de C sont proportionnellement modifiés. Ainsi, $\pi(x /_p C) = \frac{\pi(x)}{\Pi(C)}$ si $x \in C$ et 0 sinon où $/_p$ est le conditionnement basé sur le produit.

Réseaux Possibilistes (RP) Les travaux existants sur les réseaux possibilistes sont soit des adaptations directes de l'approche probabiliste [BEN 99], ou des méthodes d'apprentissage à partir de données imprécises [BOR 00]. 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. 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 des 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} : le domaine de V_i et PAR_{V_i} : l'ensemble des configurations possibles des parents de V_i. Un graphe possibiliste basé sur le produit, noté par GP_P, est un graphe possibiliste où les possibilités conditionnelles sont obtenues par le conditionnement produit. La distribution de possibilité des réseaux possibilistes basés sur le produit, notée par π_P, est obtenue par la règle de chaînage $\pi_P(V_1, .., V_N) = PROD_{i=1..N} \Pi(V_i / PAR_{V_i})$ où $PROD$ est l'opérateur produit.

3. Un modèle de RI utilisant les réseaux possibilistes

De nombreux travaux s'accordent à dire qu'il n'existe pas une définition précise de la notion de pertinence [BOR 98] [KEK 02] [BRI 03] et que cette notion est dynamique, multidimensionnelle et dépend de la perception de l'utilisateur. Nous considérons qu'il est difficile d'englober la totalité de la sémantique de la pertinence par un unique score de pertinence. Le modèle que nous proposons est capable de répondre à des propositions du type (i) il est plausible à un certain degré que le document constitue une bonne réponse à la requête, notée par $\Pi(D \mid Q)$, et, (ii) il est nécessaire, certain (dans le sens possibiliste), que le document répond à la requête, notée par $N(D \mid Q)$. Le premier type de proposition vise à éliminer certains documents de la réponse ("weak plausibility"). La seconde réponse se focalise sur les documents qui

seraient pertinents. Dans ce qui suit le modèle que nous proposons utilisant les réseaux possibilistes est détaillé.

3.1. *Architecture générale du modèle*

La topologie du réseau est représentée dans la figure (1). D'un point de vue qualitatif, le graphe permet de représenter les noeuds documents, requête, termes d'indexation et permet d'exprimer les relations de dépendance existant entre ces noeuds. Un document (D_j) est instancié ou non, prenant ses valeurs dans le domaine $\{d_j, \overline{d_j}\}$. L'activation (ou instanciation) d'un noeud document, $D_j = d_j$ (resp. $\overline{d_j}$) signifie que le document est pertinent ou non étant donnée une requête. Une requête, Q, prend ses valeurs dans le domaine $\{q, \overline{q}\}$. Nous sommes intéressés par l'instanciation de la requête, nous ne considérons que le cas $Q = q$, et nous le notons 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 l'objet (document ou requête) et donc *représentatif* de l'objet. Un terme *non-representatif*, $\overline{t_i}$, est un terme absent de la représentation de l'objet. Soit

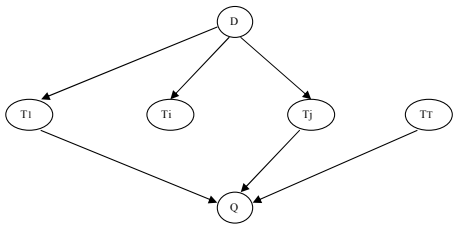

Figure 1. *Architecture générale du modèle*

$\mathcal{T}(D_j)$ (resp. $\mathcal{T}(Q)$) l'ensemble des termes indexant le document D_j (resp. la requête Q). Considérons le sous-réseau document composé des noeuds documents et de leurs termes d'indexation. Les arcs sont orientés des noeuds documents vers les noeuds termes d'indexation, exprimant ainsi les relations de dépendance existant entre les deux types de noeuds. Les termes de ce sous-réseau n'ont une existence que parce qu'ils apparaissent dans ces noeuds documents qui sont leurs parents. Considérons à présent le sous-réseau requête constitué du noeud requête et de ses termes d'indexation. La requête exprime une « demande » de documents contenant certains termes mais en excluant d'autres. La requête propage l'information aux noeuds termes qui figurent dans la collection. Ces noeuds termes forment les noeuds parents de la requête. Un terme d'indexation de la requête n'apparaissant pas dans un document donné sera considéré comme un noeud terme racine. Le système est instancié par la réception de la requête. Il existe une instanciation de l'ensemble des parents de la requête, PAR_Q, qui représente la requête dans sa forme la plus stricte (exactement telle que formulée par l'utilisateur). Soit θ^Q cette instanciation. L'ensemble des instances possibles des parents de la requête est noté θ. Nous montrerons plus tard dans cet article, comment les valeurs sont affectées aux arcs.

3.2. *Evaluation de la requête*

L'évaluation de la requête est effectuée par la propagation de l'information apportée par la requête à travers le réseau. Dans ce modèle, le processus de propagation est similaire à la propagation probabiliste Bayésienne [BEN 99] [BOR 00]. Le processus d'évaluation consiste à propager l'information *injectée* par la requête. Les noeuds reliés à la requête sont instanciés dans le but de calculer la pertinence des documents étant donnée cette requête.
Nous proposons de modéliser la pertinence par deux mesures complémentaires traduisant des aspects liés à la pertinence. Nous ne prétendons pas dans nos travaux traduire toute la sémantique liée à cette notion. Nous estimons qu'*a priori* l'utilisation de plusieurs valeurs peut traduire différents aspects de la pertinence. La **pertinence nécessaire** mesure à quel point un document doit faire partie de la liste des documents restitués. La **pertinence possible** mesure à quel point un document constitue éventuellement une réponse à une requête donnée. Le calcul de chacune de ces mesures repose sur des informations différentes. Etant donnée l'approche possibiliste choisie, nous cherchons à pouvoir restituer les documents nécessairement ou au moins possiblement pertinents étant donnée une requête. Ainsi, le processus de propagation évalue les degrés de possibilité, $\Pi(d_j \mid Q)$, et de nécessité, $N(d_j \mid Q)$, par :

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

où $\Pi(\overline{d_j} \mid Q) = \frac{\Pi(Q \wedge \overline{d_j})}{\Pi(Q)}$. La possibilité de Q est $\Pi(Q) = max(\Pi(Q \wedge d_j), \Pi(Q \wedge \overline{d_j}))$. D'après [DUB 88],[BEN 99] nous avons $\Pi(d_j \mid Q) = min(1, \frac{\Pi(Q \wedge d_j)}{\Pi(Q \wedge \overline{d_j})})$. Nous cherchons à définir $\Pi(Q \wedge D_j)$. Etant donnée la topologie du graphe, elle est de la forme :

$$\Pi(Q \wedge D_j) = \max_{\forall\, \theta^l \in \theta} (\Pi(Q \mid \theta^l) \cdot \prod_{T_i \in\, \mathcal{T}(Q) \wedge \mathcal{T}(D_j)} \Pi(\theta_i^l \mid D_j) \cdot \Pi(D_j) \cdot \prod_{T_k \in \mathcal{T}(Q) \setminus \mathcal{T}(D_j)} \Pi(\theta_k^l)) \tag{2}$$

Avec θ : les configurations possibles de l'ensemble des parents de Q, θ_i^l : l'instanciation de T_i dans la configuration θ^l ; θ^l : une configuration possible de θ.
Les configurations possibles des termes de la requête, Q, composée des termes $\{T_1, T_2\}$ sont $\theta = \{t_1 \wedge t_2;\ t_1 \wedge \overline{t_2};\ \overline{t_1} \wedge t_2;\ \overline{t_1} \wedge \overline{t_2}\}$; L'instanciation θ_1^1 du terme T_1 dans la première configuration, $\theta^1 = t_1, \wedge t_2$, est $\theta_1^1 = t_1$.
Cette quantité (2) est calculée pour $D_j \in \{d_j, \overline{d_j}\}$. 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 tête 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. Pour évaluer les documents étant donnée la requête, nous avons besoin de calculer chacun des facteurs utilisés dans (2). 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, $\Pi(Q \mid \theta)$. Les termes instanciés propagent l'information sur les documents qu'ils indexent, $\Pi(T_i \mid D_j)$. Nous définissons des postulats pour le calcul des poids des termes présents dans les documents et des termes racines, $\Pi(T_k)$. Nous montrons par la suite, le calcul de la possibilité *a priori* des documents, $\Pi(D_j)$, en absence et en présence d'information sur les documents.

3.3. *Agrégation des termes de la requête*

La possibilité de la requête étant donnée 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 déja été proposées dans les travaux de Turtle [TUR 90].
L'idée majeure de l'agrégation des termes 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 en fusionnant les fonctions de ressemblance élémentaires $\Pi(Q \mid \theta_i^l)$. Chaque $\Pi(Q \mid \theta_i^l)$ est le poids de la conformité entre l'instance θ_i^l du terme T_i avec celle de la requête (dans θ^Q). Cette configuration est en fait la configuration telle que trouvée dans un document [1]. 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. Dans notre cas, nous nous intéressons uniquement au cas $Q = q$. Une organisation possible serait de pondérer chaque terme de la requête et de calculer le poids de la jointure des termes de la requête. 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.
Conjonction, disjonction, et quantification : Pour une requête booléenne, *ET*, le processus d'évaluation restitue les documents contenant tous les termes de la requête. Ainsi, $\Pi(Q \mid \theta^l) = 1$ si $\forall\, T_i \in PAR_Q, \quad \theta_i^l = \theta_i^Q$ et 0 sinon. Chaque terme T_i parent de la requête Q doit être instancié dans θ comme dans la requête. Les documents pertinents pour ce type de requête sont les documents contenant simultanément tous ses termes. Généralement, ce type de requête est trop strict. Pour une requête booléenne, *OU*, le document est plus ou moins pertinent s'il contient au moins un terme d'indexa-

1. Nous ne considérons pas les relations de dépendance entre couples de termes ici

tion 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. Cette interprétation est trop large pour discriminer entre les documents.

Supposons 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, et que la requête contient n termes. Nous posons $f(0) = 0$ et $f(1) = 1$. Par exemple, le quantificateur $f(i/n) = 1$ si $i \geq \frac{K(\theta^l)}{n}$ et 0 sinon. D'une manière générale, f peut être une fonction non booléenne. f est un quantificateur flou [YAG 93]. L'approche quantifiée pour calculer la possibilité d'une requête Q étant donnée une configuration θ^l de tous ses parents, est donnée par $\Pi(Q \mid \theta^l) = f(\frac{K(\theta^l)}{n})$. Cette quantification, comme dans le cas d'une aggrégation disjonctive de la requête, ne permet pas de bien discriminer entre les documents de la collection.

Noisy OR : On peut supposer que les possibilités conditionnelles $\Pi(Q \mid \theta_i^l)$ ne sont pas des booléens mais dépendent d'une évaluation appropriée des termes T_i. La combinaison des termes de la requête peut être basée sur le « Noisy-Or » [PEA 88]. Cet opérateur permet de mesurer la compatibilité des instanciations des termes de la requête avec une configuration donnée. Ces termes présents dans la configuration donnée conforme à la requête sont pondérés. Pour pouvoir discriminer entre les documents, plus ce nombre de terme 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. Ce qui signifie que $\Pi(Q \mid \theta^l)$ est évaluée en termes de possibilités conditionnelles de la forme $\Pi(Q \mid t_i \wedge_{k \neq i} \overline{t_k})$ et ce en utilisant une somme probabiliste. Soit $\Pi(Q \mid t_i \wedge_{k \neq i} \overline{t_k}) = 1 - q_i$ Alors $\Pi(Q \mid \theta^l) = 0$ si $\nexists\ T_i \in PAR_Q$ tel que $\theta_i^l = \theta_i^Q$ et $\frac{1 - \prod_{i:t_i = \theta_i = \theta_i^Q} q_i}{1 - \prod_{T_k \in Par_Q} q_k}$ sinon. Uniquement, les termes instanciés positivement de la requête, $T_i = t_i$, apparaissent dans le numérateur. Le score de pertinence d'un document donné croît de manière proportionnelle au nombre de termes qu'il contient ayant la même instanciation (positive) que dans la requête [2].

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 : $\Pi(Q \mid t_i \wedge_{k \neq i} \overline{t_k}) = \frac{idf_i}{N} = nidf_i = 1 - q_i$. Un avantage majeur de ce type d'agrégation est qu'il permet d'atténuer le problème de l'explosion combinatoire liée au calcul des possibilités conditionnelles.

2. Nous supposons l'hypothèse du monde fermé ou Closed World Assumption (CWA) : $\Pi(Q \mid t_i) = \Pi(Q \mid t_i \wedge_{k \neq i} \overline{t_k})$

3.4. *Pondération des termes d'indexation*

Nous présentons dans cette section les pondérations que nous avons proposées pour les termes d'indexation. Ces pondérations sont reliées aux relations de dépendance existant entre un noeud terme et ses parents s'ils existent. En effet, lors du calcul de la pertinence d'un document étant donnée une requête, certains termes apparaissent dans le document et la requête et d'autres n'apparaissent que dans le document. [3]

Arcs document-terme $\Pi(T_i \mid D_j)$: 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 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 du document. La représentativité des termes est selon notre approche considérée sous deux angles différents mais complémentaires. 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 d'avoir la connaissance de ces deux types d'information (spécificité et/ou exhaustivité). Notre approche générale tente de distinguer les termes possiblement représentatifs des documents (ceux absents sont rejetés des représentations) de ceux qui sont nécessairement représentatifs. Ces derniers sont les termes qui suffisent à caractériser les documents.

Nous avons proposé deux approches de pondération des termes d'indexation que nous détaillons dans [BRI 05b] [BRI 05a]. Nous ne détaillons ici, que l'approche possibiliste pour la pondération des termes d'indexation. Nous traduisons la nécessaire représentativité et la plausible représentativité d'un terme basées 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 normalisée *ntf* : $\Pi(t_i/d_j) = ntf_{ij}$. Un terme de poids 0 est un terme non compatible avec le document. Si son poids vaut 1, alors le terme est possiblement représentatif du document. Un terme n'apparaissant pas dans un document est un terme non compatible avec le document et s'il apparaît avec une fréquence maximale, alors le terme est un candidat possible pour le représenter [4]. 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

3. Dans nos travaux actuels, les termes des documents absents des documents ne sont pas considérés lors des calculs de la pertinence.

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

document 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 : $\phi_{ij} = \frac{log\frac{N}{n_i}}{log(N)} \cdot ntf_{ij}$. Ce degré de nécessaire pertinence montre la nécessité qu'un terme implique un document et donc aide à restituer ce document par $N(t_i \rightarrow d_j) = \phi_{ij}$. Puisque $\Pi(\overline{d_j}) = 1$ *a priori*, alors $\Pi(t_i \mid \overline{d_j}) = \Pi(t_i \wedge \overline{d_j}) = 1 - N(t_i \rightarrow d_j) = 1 - \phi_{ij}$ et $\Pi(\overline{t_i} \mid \overline{d_j}) = 1$
Termes racines : Les termes racines sont les termes qui apparaîssent dans la requête mais pas dans le document. Lors du processus de propagation ces termes sont instanciés par la requête et notre modèle mesure l'absence de ces termes. Dans notre approche, un terme discriminant absent du document pénalise la pertinence de ce document. Nous avons présenté dans [BRI 05a] un nouveau facteur discriminant, noté ndf_i, utilisant l'entropie de Shanon. Ce facteur utilise la densité d'un terme t_i dans un document d_j que nous mesurons par $\frac{tf_{ij}}{l_j}$ où $l_j = \sum_j tf_{ij}$. Le facteur ndf_i est obtenu par $ndf_i = \frac{-\sum_{t_i} p_{ij} log(p_{ij})}{max_{k \in T} ndf_k}$, avec p_{ij} proportionnel à $\frac{tf_{ij}}{l_j}$ et T le nombre de termes de la collection. Ainsi, $\forall T_i \notin \mathcal{T}(D_j),\ \Pi(\theta_i) = 1$ si $\theta_i^Q = \overline{t_i}$ et ndf_i sinon. Un terme peu densément distribué dans la collection minimise le facteur ndf et inversement.

3.5. *Possibilité a priori des documents*

En absence d'information, la possibilité *a priori* d'un noeud document est uniforme $\Pi(d_j) = \Pi(\overline{d_j}) = 1$. Nous pouvons obtenir des connaissances sur les documents étant donnée l'importance de ses termes, sa longueur etc. Si nous sommes intéressés par les documents longs, la possibilité *a priori* d'un document instancié à $D_j = d_j$ devient : $\Pi(d_j) = \frac{l_j}{max_{k=1,..,N} l_k} = nl_j$ où 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. De plus, $\Pi(\overline{d_j}) = 1$.

4. 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. Parmi ces requêtes, 6 d'entre elles contiennent des termes qui ne figurent dans aucun document de la collection. Ces requêtes ne sont pas évaluées par notre système.
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. Les para-

mètres dans notre système représentent les informations considérées lors du processus de propagation déclenchée par la requête (formule 2).

Le modèle optimal Nous décrivons dans cette section les instanciations prises par les paramètres du modèle optimal, qui a permis d'obtenir les meilleurs performances. Les paramètres ont été fixés pour ce modèle tels que décrits dans le tableau 1. Dans

Tableau 1. *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

$\Pi(Q \mid T_i)$	t_i	$\overline{t_i}$
Q	$nidf_i$	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

le tableau 1 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

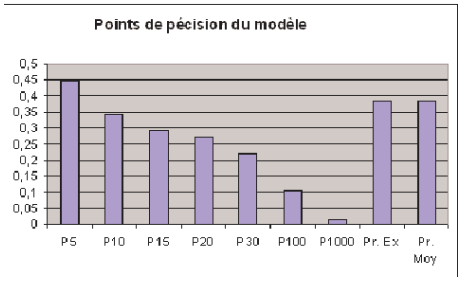

Figure 2. *Points de précision du modèle optimal*

conditionnelle ($\Pi(T_i \mid D_j)$). Les termes de la requête sont agrégés par l'opérateur du *Noisy Or*. La figure 2 présente les valeurs des points de précision obtenues pour les 34 requêtes évaluées. La précisions exacte et moyenne de ce modèle de base sont de 0.3661 et 0.3821 respectivement. Nous remarquons dans la figure 2 que l'écart entre les points de précision P_5 et P_{10} 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 ».

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 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 $OKAPI$. La

pondération des termes utilisés est $BM-25$ [ROB 94]. 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 remar-

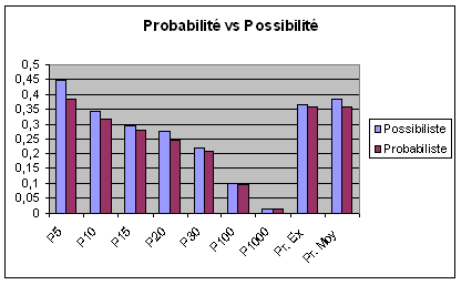

Figure 3. *Comparatif des deux systèmes : possibiliste et probabiliste*

quons 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 $P_5, .., P_{20}$ obtenues par notre système sont plus élevées. Nous obtenons une amélioration de plus de 14% pour la précision à 5 (P_5). D'une manière générale, comme présenté dans le tableau 2, les précisions $P_5, .., P_{20}$ obtenues par l'utilisation de notre approche sont supérieures de plus de 5% au modèle $OKAPI$. $P_i^{Possibiliste}$

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

et $P_i^{Probabiliste}$ désignent la précision au point P_i 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 34 requêtes évaluées par les 2 systèmes, le système possibiliste améliore les précisions à 5 (P_5) de 14 d'entre elles, et obtient les mêmes valeurs pour 13 d'entre elles. Le système $OKAPI$ obtient de meilleures valeurs P_5 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.

5. Conclusion 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 de la notion 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.

6. Bibliographie

[BEN 99] BENFERHAT S., DUBOIS D., L.GARCIA, PRADE H., « Possibilistic logic bases and possibilistic graphs », *Proc. of Conference on Uncertainty in AI*, 1999, p. 57-64.

[BOR 98] BORLUND P., INGWERSEN P., « Measures of relative relevance and ranked half-life : performance indicators for interactive IR », *Proc. of the ACM-SIGIR Conference*, 1998, p. 24-28.

[BOR 00] BORGELT C., GEBHARDT J., KRUSE R., « Possibilistic graphical models », *Computational Intelligence in Data Mining, CISM Courses and Lectures, Springer*, vol. 408, 2000, p. 51-68.

[BRI 03] BRINI A., BOUGHANEM M., « Relevance feedback : introduction of partial assessments for query expansion », *Proc. of the Conference (EUSFLAT), Zittau*, 2003, p. 67-72.

[BRI 05a] BRINI A., CAMPOS L., DUBOIS D., BOUGHANEM M., « Query Propagation in Possibilistic Information Retrieval Networks », *Proc. of (EUSFLAT 2005), Barcelona*, 2005, page CD.

[BRI 05b] BRINI A. H., « Un modèle de recherche d'information basé sur les réseaux possibilsites », Thèse de doctorat, Université de Toulouse III, UPS, 2005.

[DUB 88] DUBOIS D., PRADE H., *Possibility Theory*, Plenum, 1988.

[KEK 02] KEKÄLÄINEN J., JÄRVELIN K., « Evaluating information retrieval systems under the challenges of interaction and multidimensional dynamic relevance. », *Emerging Frameworks and Methods*, 2002, p. 253-270.

[PEA 88] PEARL J., *Probabilistic Reasoning in Intelligent Systems*, Morgan Kaufmann, 1988.

[ROB 94] ROBERTSON S., WALKER S., « Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval », *Proc. of the ACM-SIGIR Conference*, 1994, p. 232-241.

[TUR 90] TURTLE H., CROFT W., « Inference networks for document retrieval », *Proc. of the ACM-SIGIR Conference*, 1990, p. 1-24.

[YAG 93] YAGER R. R., LARSEN H. L., « Retrieving information by fuzzification of queries », *Journal of Intelligent Information Systems*, vol. 2, n° 4, 1993, p. 106-119.

[ZAD 78] ZADEH L. A., « Fuzzy Sets as a Basis for a theory of Possibility », *Fuzzy Sets and Systems*, vol. 1, 1978, p. 3-28.