Recherche d'Information Structurée

Vers un modèle possibiliste pour la recherche d'information dans des documents structurés

Fatma-Zohra BESSAI MECHMACHE[(1)] et Mohand BOUGHANEM[(2)]

(1) Ministère de l'Enseignement Supérieur et de la Recherche Scientifique,
Centre de Recherche sur l'Information Scientifique et Technique (CERIST),
Rue des 3 Frères Aissou, Alger, Algérie
zbessai@cerist.dz

(2) Université Paul Sabatier
IRIT-SIG
118 Route de Narbonne 31062
bougha@irit.fr

Chercheur

Résumé Dans cet article, nous nous intéressons à la Recherche d'Information (RI) dans des documents structurés en XML. Pour cela, nous présentons un modèle pour la recherche d'information structurée, basé sur les réseaux possibilistes. Les relations document-éléments et éléments-termes sont modélisées par des mesures de possibilité et de nécessité. Dans ce modèle, la requête de l'utilisateur déclenche un processus de propagation pour retrouver des documents ou des portions de documents nécessairement ou au moins possiblement pertinents par rapport à la requête. Un exemple d'une telle recherche est proposé de façon à illustrer l'approche présentée.

Abstract In this paper, we are interested in Information Retrieval in structured document in XML. For this, we present a model for the structured information retrieval, based on the possibilistic networks. The document - elements and elements - terms relations are modelled by measures of possibility and necessity. In this model, the user's request starts a process of propagation to recover the documents or the portions of documents necessarily or at least possibly relevant. An example of such a research is proposed in order to illustrate the presented approach.

Mots-clés: Recherche d'information structurée, documents XML, théorie des possibilités, réseaux possibilistes, indexation.

Keywords: Structured information retrieval, XML document, possibilistic theory, possibilistic network.

1 Introduction

Le développement du document électronique et du Web ont vu émerger puis s'imposer des formats de données structurés, tels que le SGML (Standard Generalized Markup Language) et le XML (eXtensible Markup Language), permettant de représenter l'information sous une forme plus riche que le simple contenu et adaptée à des besoins spécifiques. Ces nouveaux formats permettent de représenter conjointement l'information textuelle et l'information de structure d'un document. La connaissance de la structure des documents est une ressource additionnelle qui devrait être exploitée pendant la recherche d'information afin de mieux exprimer un besoin d'information. Dans le contexte de la RI, la question majeure soulevée par ce type de document concerne la manière de manipuler efficacement la structure et le contenu du document pour mieux répondre aux besoins de l'utilisateur. Ces besoins peuvent être formulés par le biais de requêtes formées que de mots clé ou par des requêtes comportant des mots clés et des contraintes structurelles (des balises).

Les approches proposées dans ce cadre peuvent être classées en deux principales catégories : (i) l'***approche orientée données*** utilise des techniques développées par la communauté des Bases de Données (BD), (ii) l'***approche orientée documents*** est prise en charge par la communauté Recherche d'Information (RI). Les approches orientées BD s'intéressent davantage à la structure du document. Plusieurs langages ont été définis [Buneman 96], Lorel [Abiteboul 97], XML-QL [Levy 98], XQL [Chamberlin 00], XML-GL [Ceri 99]. Les ***approches orientées documents*** considèrent les documents XML comme une collection de documents textes comportant des balises (éléments) et des relations entre ces balises. Les balises sont utilisées comme moyen pour mieux identifier la pertinence d'une partie de document vis –à- vis d'une autre partie. La majorité des travaux ont en fait adapté les modèles de RI reconnus pour traiter les documents XML [Lalmas 97], [Hayashi 00], [Schlieder 02], [Fuhr 01], [Piwowarski 02], [Abolhassani 04], [Ogilvie 03], [Sauvagnat 05].

Si les approches orientées BD permettent de traiter efficacement la structure des documents XML, elles sont cependant limitées pour le traitement de la partie textuelle des documents. Les mots clés sont en effet traités de façon binaire (présent /absent), or il a été démontré en RI textuelle [Salton 83] que la prise en compte des poids des mots-clés dans un document est primordiale, voire nécessaire. Ceci permet de mesurer un degré de pertinence d'un document (ou d'une partie de document) vis-à-vis d'une requête et donc de renvoyer à l'utilisateur une liste triée de résultats, comme le proposent les approches de RI.

Un second problème auquel sont confrontés les systèmes d'accès aux documents structurés est celui de l'unité d'information à sélectionner dans le cas ou la requête de l'utilisateur ne comporte que des mots clés. En effet, les techniques classiques de RI (plein texte) considèrent le document entier comme un granule d'information

indivisible, or dans le cas des documents XML tout élément (sous-arbre d'un document XML) peut être une réponse potentielle à la requête de l'utilisateur. Le défi à relever est alors d'arriver à identifier automatiquement l'unité d'information, en l'occurrence les parties du document XML, répondant à la fois de manière exhaustive et spécifique [Sauvagnat 05] à la requête de l'utilisateur.

Notre objectif est de proposer un modèle permettant de sélectionner automatiquement l'élément (ou l'ensemble d'éléments) du document qui répond le mieux au besoin de l'utilisateur formulé à travers une liste de mots clés. De plus, afin de mieux prendre en compte la structure hiérarchique d'un document XML, les dépendances entre des éléments, ainsi que l'imprécision liée à la notion de pertinence, le modèle que nous proposons trouve ses fondements théoriques dans les réseaux possibilistes. La structure réseau fournit une manière naturelle de représenter les liens entre, un document, ses éléments (balises) et son contenu. Quant à la théorie des possibilités, elle permet de quantifier de manière qualitative et quantitative les différents liens sous jacents. Elle permet notamment d'exprimer le fait qu'un terme est certainement ou possiblement pertinent vis-à-vis d'un élément et/ou d'un document et de mesurer à quel point un élément (ou un ensemble d'éléments) peut nécessairement ou possiblement répondre à la requête de l'utilisateur.

Cet article est structuré de la manière suivante. La section 2 présente un état de l'art succinct sur la RI structurée. La section 3 décrit quelques concepts de base de la théorie des possibilités. La section 4 est consacrée à la description du modèle que nous proposons. Nous montrons, en section 5 un exemple illustrant ce modèle.

2 Etat de l'art et motivation

Les approches proposées pour adresser le problème de requêtes composées que de mots clés peuvent être divisées en 2 catégories principales [Abolhassani 04]. Ces approches se distinguent par leur manière d'indexer les éléments d'un document XML.

La première catégorie considère que n'importe quel élément (ou sous-arbre d'éléments) XML peut potentiellement être renvoyé à l'utilisateur. Chaque élément est donc traité comme une unité atomique. Le texte de chaque élément est vu comme l'agrégation du texte contenu dans ses descendants. L'élément composé de la racine, correspond au document entier. Les plus petits sous arbres sont ceux qui comportent seulement une feuille de l'arbre. Afin d'identifier les éléments à renvoyer en réponse à la requête, un score de pertinence est calculé entre la requête et chacun des sous arbres. Les sous-arbres résultats sont alors triés et renvoyés à l'utilisateur selon leurs scores de pertinence. Des exemples de ces approches peuvent être trouvés dans [Sigurbjornsson 03], [Abolhassani 04], basé sur les modèles de langue.

La seconde catégorie considère un document XML comme un ensemble d'unités disjointes. En fait une unité est composée d'un chemin qui va de la racine vers une feuille de l'arbre. Chaque élément d'un document est vu comme l'union d'une ou

plusieurs de ces unités. De la même façon que les approches précédentes, tout nœud (sous-arbre) du document peut être renvoyé à l'utilisateur, mais le score de pertinence d'un élément donné vis-à-vis d'une requête est calculé en agrégeant les scores de pertinence des unités disjointes qui le composent. Plusieurs modèles ont été proposés dans ce cadre. On y trouve notamment, une approche basée sur les modèles de langue définie dans [Ogilvie 03]. Plus précisément, deux modèles de langages sont proposés, un modèle pour les unités disjointes (unitaires) et un modèle pour les unités composées (unité de haut niveau). Le modèle des nœuds de haut niveau (composés) est en fait la somme des modèles de langue de ses unités (composées ou unitaires). Une méthode d'augmentation est également proposée dans [Govert 02]. Les nœuds disjoints considérés dans cette approche ne sont pas nécessairement des nœuds feuilles. En fait les nœuds indexables, donc ceux que l'on peut renvoyer à l'utilisateur, sont définis au préalable. De ce fait, les termes des nœuds feuilles, non indexables, sont propagés jusqu'au noeud indexable le plus proche. Afin de préserver les unités disjointes, on ne peut associer à un noeud que des termes reliés à ses noeuds descendants. Chaque nœud indexable de l'arbre calcule alors un score de pertinence vis-à-vis de la requête. Ce score est calculé grâce à la propagation des poids des termes les plus spécifiques dans l'arbre du document. Les poids sont cependant diminués par multiplication par un facteur, nommé facteur d'augmentation. Les nœuds renvoyés sont ordonnés selon leurs degrés de pertinence. On trouve également dans [Sauvagnat 05] le même principe de propagation. Mais contrairement au modèle de Fuhr, Saugnavat et. al ne définissent pas au préalable les unités indéxables et ils propagent les scores de pertinence calculés au niveau des feuilles vers les nœuds de plus haut niveau.

D'autres approches proches du modèle que nous proposons ont également été proposées. Nous citons par exemple celle basée sur la théorie de l'évidence de Dempster-Schafer [Lalmas 97], [Lalmas 04]. Pour tout document et toute requête, il est possible de définir deux mesures : la croyance et l'incertitude. Un opérateur d'agrégation permet de combiner la croyance de ces sous structures pour calculer la pertinence d'un document. Nous citons aussi l'approche de Piwowarski [Piwowarski 02] qui s'appuie sur le formalisme des Réseaux Bayésiens (RB) et qui calcule des probabilités conditionnelles qui sont l'entrée du RB, en fonction de la question et la recherche correspond à l'inférence dans le RB.

Notre approche rentre dans la seconde catégorie des approches. Elle est basée sur la théorie possibiliste [Borgelt 00], et plus particulièrement les réseaux possibilistes. Ces réseaux offrent un modèle simple et naturel pour à la fois représenter la structure hiérarchique des documents XML et pour manipuler l'information incertaine inhérente à la recherche d'information de manière générale. On trouve cette incertitude dans, la notion de pertinence d'un document vis-à-vis d'une requête, le degré de représentativité d'un terme dans un document ou une partie de documents et l'identification de la partie pertinente répondant à la requête. Dans ce cadre, afin d'identifier la partie pertinente qui répond à la requête, contrairement aux approches proposées dans la littérature, qui sélectionnent comme nous l'avons vu, le sous arbre susceptible d'être pertinent, notre approche permet d'identifier et de sélectionner, de

manière naturelle, l'élément (ou l'ensemble d'éléments) du document XML susceptible de répondre à la requête. De plus, nous interprétons la notion de pertinence par deux dimensions :

- une dimension qui mesure à quel point il est certain qu'une « composition d'éléments d'un document » est pertinente vis-à-vis de la requête
- une dimension qui mesure à quel point il est possible qu'une « composition d'éléments d'un document» est possiblement pertinente pour la requête.

Outre les points cités ci-dessus, le cadre théorique qui supporte nos propositions, en l'occurrence les réseaux possibilistes nous différencient clairement des cadres utilisés dans les approches précédentes.

3 La théorie des possibilités

La théorie des possibilités a été introduite en 1978 par Lotfi A. Zadeh [Zadeh 78], en liaison avec la théorie des sous-ensembles flous, pour permettre de raisonner sur des connaissances imprécises. La notion de base de la théorie des possibilités est la distribution de possibilité, notée Π, qui est une fonction de l'ensemble Ω vers une échelle totalement ordonnée qui a une valeur maximale et une valeur minimale. Généralement, cette échelle est l'intervalle unitaire [0, 1].

Intuitivement, une distribution de possibilité peut représenter soit les croyances d'un agent ou bien ses préférences. $\Pi(w)$ représente le degré de compatibilité de l'interprétation w avec les croyances disponibles sur le monde réel si l'on représente des informations incertaines (ou le degré de satisfaction de w si l'on représente des préférences).

Par convention, $\Pi(w)=1$ signifie qu'il est totalement possible que w soit le monde réel (ou que w est totalement satisfaisante), $1>\Pi(w)>0$ signifie que w est quelque peu possible (ou satisfaisante), et enfin $\Pi(w)=0$ signifie que w n'est certainement pas le monde réel (ou est totalement insatisfaisante).

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 $\prod(A)$, et la mesure de nécessité N(A). La possibilité d'un événement A, notée $\prod(A)$ est obtenue par $\prod(A) = \max_{x \in A} \pi(x)$ et décrit la situation la plus normale dans laquelle A est vraie.

La nécessité $N(A) = \min_{x \notin A} 1 - \pi(x) = 1 - \prod(\neg A)$ d'un événement A reflète la situation la plus normale dans laquelle A est faux. La distance entre N(A) et $\prod(A)$ 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 (ensemble d'états possibles), 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, $\pi(x /p\ C) = \pi(x)/ \prod(C)$ si $x \in C$ et 0 sinon où /p est le conditionnement basé sur le produit.

Réseaux Possibilistes (RP) Un graphe possibiliste orienté sur un ensemble de variables $V = \{V_1, V_2, \ldots, V_N\}$ 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 entre ces variables. 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 $\max_{vi} \prod(v_i)=1$, $\forall v_i \in dom(V_i)$. (ii) Si V_i n'est pas un noeud racine, la distribution conditionnelle de V_i dans le contexte de ses parents doit satisfaire $\max_{vi} \prod(vi/PAR_{Vi}) = 1$, $\forall vi \in dom(V_i)$ où $dom(V_i)$: le domaine de V_i et PAR_{Vi} : 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 $\pi_P (V_1, .., V_N) = PROD_{i=1..N} \prod(V_i/PAR_{Vi})$ où PROD est l'opérateur produit.

4 Architecture du modèle possibiliste pour la RI structurée

Le modèle que nous proposons est représenté par un réseau possibiliste d'architecture illustrée par la figure (1). D'un point de vue qualitatif le graphe permet de représenter les nœuds documents, termes d'indexation, nœuds (balises d'un document XML). Les liens entre les nœuds permettent de représenter les relations de dépendances entre les différents nœuds.

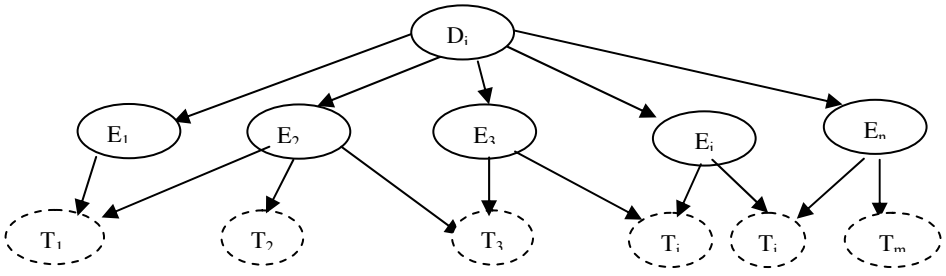

Figure1*: Architecture du modèle*

4.1 Description du modèle

Les nœuds documents représentent les documents de la collection. Chaque noeud document D_i, représente une variable aléatoire binaire prenant des valeurs dans l'ensemble $dom(D_i) = \{d_i, \neg d_i\}$, où d_i représente 'le document D_i est pertinent pour une requête donnée', et $\neg d_i$ représente 'le document D_i n'est pas pertinent pour la requête donnée'.

Les nœuds E_1, E_2, …E_n, représentent les balises du document D_i. Chaque noeud E_i, représente une variable aléatoire binaire prenant des valeurs dans l'ensemble $dom(E_i) = \{e_i, \neg e_i\}$. L'instanciation $E_i = e_i$ signifie que l'élément 'e_i' est pertinent pour la requête; $E_i = \neg e_i$ signifie que l'élément 'e_i' est non pertinent pour la requête.

Les nœuds $T_1, T_2, \ldots, T_m$ sont les nœuds termes. Chaque noeud terme Ti représente une variable aléatoire binaire prenant des valeurs dans l'ensemble $dom(T_i) = \{t_i, \neg t_i,\}$ où L'instanciation $T_i = t_i$ signifie que le terme 'T_i' est représentatif du nœud père auquel il est rattaché, $T_i = \neg t_i$ signifie que le terme 'T_i' est non représentatif du nœud père auquel il est relié. Il faut noter qu'un terme est relié aussi bien au nœud qui le comporte ainsi qu'à tous les ascendants de ce dernier

Le passage du document vers la représentation sous forme de réseau possibiliste se fait de manière assez simple. Il consiste à ramener tous les nœuds (balises du document) au niveau des variables E_i. Les valeurs qui seront assignées aux arcs de dépendances entre nœuds termes - nœuds balises et nœuds balises - noeud document dépendent du sens que l'on donne à ces liens.

Chaque variable structurelle E_i, $E_i \in E = \{E_1, E_2\ldots, E_m\}$, dépend directement de son nœud parent qui est le nœud racine D, dans le réseau possibiliste du document.

Chaque variable de contenu T_i, $T_i \in T = \{T_i, \ldots, T_k\}$ dépend uniquement de sa variable structurelle (élément structurel ou balise). Il faut également noter que la représentation fait apparaître un seul document. En fait on considère que les documents sont indépendants les uns des autres donc on peut raisonner en considérant le sous réseau qui représente le document que l'on traite.

Nous notons par T(E) (resp. T (Q)) l'ensemble des termes d'indexation des éléments du document (resp. de la requête).

Les arcs sont orientés et on en distingue deux types :
– *Les liens termes-balises*. Ces liens relient chaque noeud terme $T_i \in T(E)$ à chaque nœud E_i ou il apparaît.
– Les liens balises-document. Ces liens relient chaque nœud balise de *E* au document qui le comporte, en l'occurrence *D* dans notre cas. Nous discuterons dans la section 4.3, l'interprétation que nous donnons à ces différents liens et la manière de les quantifier.

4.2 Evaluation d'une requête par propagation

L'évaluation d'une requête est effectuée par à un processus de propagation [Benferhat 99] [Borgelt 00]. Elle consiste à injecter une nouvelle évidence à travers les arcs activés du réseau pour rechercher les documents et les éléments pertinents par rapport à la requête. Comme nous l'avons souligné précédemment, nous modélisons la pertinence selon deux dimensions : la nécessité et la possibilité de pertinence. Notre modèle doit être capable d'inférer des propositions du type :

- « le document d_i est pertinent pour la requête Q » est possible à un certain degré ou non, quantifiée par $\prod(d_i/Q)$.

- « le document di est pertinent pour la requête Q » est certain ou non, quantifiée par $N(d_i/Q)$.

Le premier type de proposition permet d'éliminer les documents (et les éléments) non pertinents, c'est-à-dire ceux qui ont une faible possibilité. La deuxième proposition concentre l'attention sur ceux qui semblent très pertinents.

Pour le modèle de base présenté ici, nous prendrons les hypothèses suivantes :

Hypothèse1 : Un document a autant de possibilité d'être pertinent que non pertinent pour un utilisateur donné, soit $\prod(d_i) = \prod(\neg d_i) = 1$, quelque soit i.

Hypothèse2 : La requête est composée d'une simple liste de mots-clés (interprétée de manière conjonctive) $Q= \{t_1 \wedge t_2 \wedge .. \wedge t_l\}$. L'importance relative des termes entre eux dans la requête est ignorée.

Afin d'évaluer les quantités, $\prod(d_i/Q)$ *et* $N(d_i/Q)$, nous procédons de la manière suivante(NB : nous nous sommes fortement inspirés du modèle développé par A. Brini, M. Boughanem et D . Dubois [brini 05] et utilisé pour la recherche plein texte,

$$\Pi(d_i / Q) = \frac{\Pi(Q \wedge d_i)}{\Pi(Q)} \quad et \quad N(d_i / Q) = 1 - \Pi(\neg d_i / Q)$$

En considérant la quantité $\Pi(Q)$ indépendante des documents, donc le classement obtenus par $\Pi(d_i / Q)$ sera identique à celui que l'on obtiendrait avec $\Pi(Q \wedge d_i)$. De ce fait, compte tenu de l'hypothèse 2, et étant donné l'architecture du modèle (voir figure1), le facteur $\Pi(Q \wedge d_i)$ se quantifie comme suit :

$$\Pi(Q \wedge d_i) = \max_{\theta^e \in \theta^E} (\Pi(Q/\theta^e) * \Pi(\theta^e/d_i) * \Pi(d_i)) \quad [1]$$

Ou θ^e représente une instanciation possible des variables balises, c'est-à-dire les parents des termes de la requête parmi toutes les configurations possibles définies par θ^E. A titre d'exemple, considérons que les termes de la requête sont reliés à deux balises (e_1 et e_2), les instanciations possibles sont $\{e_1 \wedge e_2 , \neg e_1 \wedge e_2, e_1 \wedge \neg e_2, \neg e_1 \wedge \neg e_2\}$.

De plus en supposant que les variables termes et les variables balises sont indépendantes. L'équation [1] devient alors :

$$\Pi(Q \wedge d_i) = \max_{\forall \theta^e \in \theta^E} (\Pr od(\Pr od_{E_j \in \theta^e} (\Pi(t_i / \theta_j^e))) * \Pr od_{E_j \in \theta^e} (\Pi(\theta_j^e / d_i)) * \Pi(d_i))$$

[2]

Où *Prod* : signifie produit (nous avons utilisé ce symbole au lieu de $\prod$ pour ne pas le confondre avec le symbole désignant la possibilité).

$t_i \in T(E) \wedge T(Q)$: représente les termes de la requêtes qui indexent les balises.

θ_j^e : représente l'instance de E_j dans la configuration θ^e (exemple la valeur de E1 dans la configuration $\{e_1 \wedge e_2\}$ est e_1)

A l'issue de la propagation de la requête chaque document avec une certaine configuration (instances de balises) aura une valeur de nécessité et une valeur de possibilité. La liste (document configuration) est triée par ordre de nécessité puis par possibilité de pertinence. Nous rappelons, comme nous l'avons souligné précédemment, que la sélection des parties pertinentes (unités d'information) est sous-jacente au modèle. En effet, la formule (2) calcule la pertinence en considérant toutes les combinaisons possibles des balises. Le facteur θ^e donne une instanciation possible des balises. La combinaison de balises qui sera sélectionnée sera celle qui comporte obligatoirement les termes de la requête (à cause du produit) et qui présente la meilleure pertinence (la pertinence maximale) en termes de nécessité et/ou de possibilité.

Une étape primordiale de notre modèle concerne la quantification et le sens que nous donnons aux différents arcs. Afin d'évaluer les différents degrés Π et N entre les nœuds du réseau, nous nous sommes inspirés du modèle [Brini 05] utilisé pour la recherche d'information plein texte.

4.3 Détermination de la valeur des arcs

4.3.1 Valeur de l'arc nœud balise- nœud terme (pertinence du contenu)

Dans le domaine de la recherche d'information, les termes utilisés pour représenter le contenu d'un document, sont pondérés de manière à mieux caractériser le contenu de ce document [Baeza 99]. Le même principe est utilisé en recherche d'information structurée. Les poids sont généralement calculés en utilisant l'information de fréquence d'apparition des termes et la fréquence inverse du document.

En recherche d'information, il a été montré [Rölleke 02], [sauvagnat 2005] que les performances du système peuvent être améliorées si l'on représente un élément par son propre contenu et par celui de ses éléments fils. Dans notre modèle, nous distinguons les termes possiblement représentatifs des éléments du document et ceux

nécessairement représentatifs de ces éléments (termes qui suffisent pour caractériser les éléments). Pour ce faire, nous considérons les hypothèses suivantes :

- Hypothèse3 : Un terme est représentatif d'un élément (balise) s'il apparaît dans cet élément.

- Hypothèse4 : Un terme est nécessairement représentatif de l'élément s'il apparaît fréquemment dans cet élément et peu fréquemment dans les autres éléments du document.

D'après l'hypothèse3, la possibilité de pertinence d'un terme (t_i) pour représenter un élément (e_j), notée $\Pi(t_i / e_j)$, est calculée comme suit :

$$\Pi(t_i / e_j) = tf_{ij} / \max_{\forall t_k \in e_j} (tf_{kj})$$

Un terme ayant un degré de possibilité 0 signifie que le terme n'est pas représentatif de l'élément. Si le degré de possibilité est strictement supérieur à 0, alors le terme est possiblement représentatif de la balise. S'il apparaît avec un degré de possibilité maximum, alors il est considéré comme le meilleur candidat potentiel pour la représentation et donc la restitution de la balise.

Notons que $max(\Pi(t_i / e_j)) = 1$, quelque soit $t_i \in e_j$

Dans un document structuré, un terme nécessairement représentatif d'un élément est un terme qui contribue à sa restitution en réponse à une requête. Ce terme est appelé terme discriminant et c'est un terme qui apparaît fréquemment dans peu d'éléments du document structuré [Brini 05]. Le facteur communément utilisé en RI pour quantifier le pouvoir discriminant d'un terme est *idf* (*ief* en RI structurée). Par conséquent, un degré de nécessaire pertinence, β_{ij}, du terme ti pour représenter l'élément e_j, sera défini par :

$$N(t_i \rightarrow e_j) \geq \beta_{ij} = \mu(tf_{ij} * ief_{ij}) * idf = \mu(tf_{ij} * \log(\frac{Ne}{ne_i}) * \log(\frac{N}{n_i}))$$

Avec, N et Ne_i : respectivement le nombre de document et d'éléments dans la collection ; n_i et ne_i : respectivement le nombre le document et le nombre d'éléments contenant le terme t_i et μ : est une fonction de normalisation. Une manière simple de normaliser est de diviser par la valeur maximale du facteur.

Il faut noter que cette formule a été choisie compte tenu des performances qu'elle a obtenues dans le cadre des travaux de Sauvagnat et Boughanem [sauvagnat 05].

Ce degré de nécessaire pertinence va permettre de borner la possibilité que le terme est compatible avec le rejet de l'élément par :

$\Pi(t_i / \neg e_j) \leq 1 - \beta_{ij}$ (ceci est déduit par définition dans l'approche possibiliste)

Nous récapitulons la distribution de possibilité définie sur le produit cartésien $\{ej, \neg ej\} \times \{t_i, \neg t_i\}$ par le tableau suivant :

Π	e_j	$\neg e_j$
t_i	$tf_{ij} / max(tf_{kj}), (\forall t_k \in e_j)$	$1 - \beta_{ij}$
$\neg t_i$	1	1

Tableau 1. *Distribution de possibilité définie sur l'ensemble des termes T*

4.3.2 Valeur de l'arc nœud document – nœud balise (propagation de pertinence)

L'arc nœud document – nœud balise (ou arc racine-élément) indique l'intérêt de propager une information d'un élément vers le nœud racine document. Les noeuds apparaissant près de la racine d'un arbre paraissent plus porteurs d'information pour le noeud racine que ceux situés plus bas dans l'arbre [Sauvagnat 05]. Il semble ainsi intuitif que plus grande est la distance entre un noeud et la racine de l'arbre, moins il contribue à sa pertinence. Nous modélisons cette intuition par l'utilisation dans la fonction de propagation du paramètre *dist(racine, n)*, qui représente la distance entre le noeud racine et un de ses noeuds descendants (balises) n dans l'arbre hiérarchique du document, c'est à dire le nombre d'arcs séparant les deux noeuds.

Le degré de possibilité de propagation d'une balise (e_i) vers le nœud document di est défini par $\prod (e_j / d_i)$ et est quantifié comme suit :

$$\prod (e_j / d_i) = \alpha^{dist(di,\, ej)-1}$$

Avec :

- $dist(d_i, e_j)$ la distance de la balise e_j à la racine di conformément à la structure hiérarchique du document.
- $\alpha \in]0..1]$ est un paramètre permettant de quantifier l'importance de la distance séparant les nœuds balises (éléments structurels du document) à la racine du document.

Concernant la nécessité de propager, de manière intuitive, on peut penser que le concepteur d'un document utilise les noeuds de petite taille pour faire ressortir des informations importantes. Ces noeuds peuvent ainsi donner des indications précieuses sur la pertinence de leurs noeuds ancêtres. Un noeud titre dans une section par exemple permet de situer avec précision le sujet de son noeud ancêtre section. Il est donc nécessaire de propager le signal calculé au niveau du nœud vers le nœud racine. Pour répondre à cette intuition, nous proposons de calculer la nécessité de propagation de pertinence d'un élément e_j vers le nœud racine di, notée $N(e_j \rightarrow d_i)$ comme suit :

$$N(e_j \rightarrow d_i) = 1 - \frac{le_j}{dl}$$

Soit le_j la taille du nœud balise (élément structurel) ej et dl la taille d'un document (en nombre de termes). D'après la formule, plus un terme est de taille petite plus la nécessité de le propager est grande.

Par conséquent, $\prod (e_j / \neg d_i) = le_j/dl$

Nous récapitulons la distribution de possibilité définie sur le produit cartésien {di, ¬ di} x {ej, ¬ ej} par le tableau suivant :

	d_i	$\neg d_i$
e_i	$\alpha^{dist(d_i, e_j)-1}$	le_j/dl
$\neg e_i$	1	1

Tableau 2. *Distribution de possibilité définie sur l'ensemble des éléments E*

5 Exemple illustratif

Un exemple de document XML (un extrait d'un document) relatif à un ouvrage va être utilisé pour illustrer notre approche. Le document XML exemple ainsi que le réseau possibiliste qui lui est associé sont présentés dans ce qui suit :

```
<Ouvrage>
        <Titre > Recherche d'Information </Titre >
        <Résumé>Devant la masse croissante d'information …</Résumé>
….
        <Chapitre>
        <Titre chapitre> Indexation </titre chapitre>
        <Paragraphe> L'indexation est le processus destiné à représenter par les
              éléments d'un langage documentaire ou naturel des … </Paragraphe>
        </Chapitre>
</Ouvrage>
```

La structure hiérarchique du document XML 'ouvrage' est comme suit :

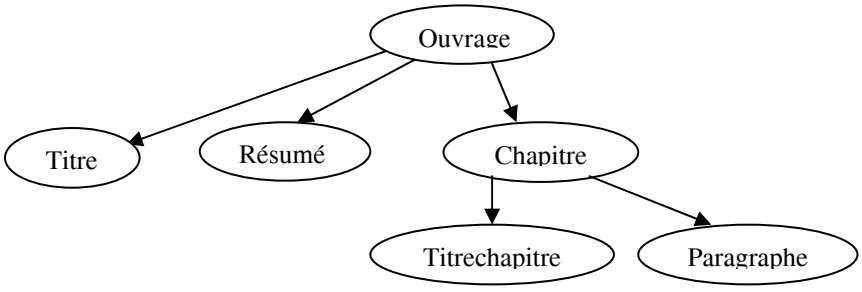

Figure 3. *Structure hiérarchique du document XML 'ouvrage'*

Le réseau possibiliste associé au document XML 'ouvrage' est comme suit :

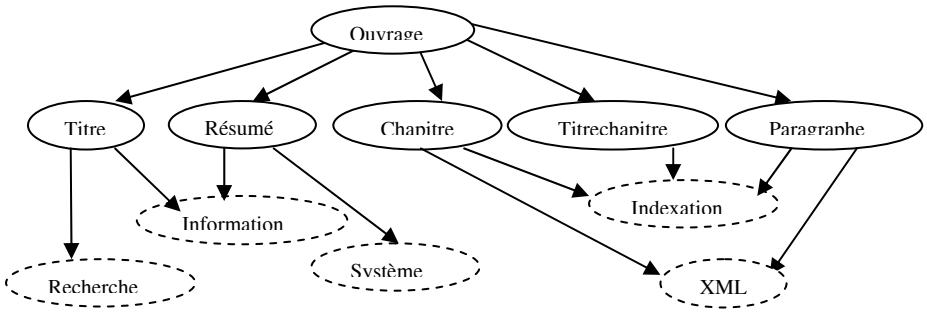

Figure 4. *Réseau possibiliste du document XML 'ouvrage'*

Pour cet exemple, l'ensemble des éléments E= {e_1=Titre, e_2=Résumé, e_3=Chapitre, e_4=Titrechapitre, e_5=Paragraphe}. L'ensemble des termes d'indexation des éléments, calculé en utilisant le contenu de chaque élément ainsi que celui de ses éléments fils dans le document, est tel que T(E) = {t_1=Recherche, t_2=Information, t_3=Système, t_4=Indexation, t_5=XML}. On ne considère que quelques termes pour ne pas encombrer l'exemple.

La matrice contenant les valeurs des arcs nœud balise - nœud terme du réseau possibiliste du document 'ouvrage' est donnée dans le tableau 3. Nous rappelons qu'un terme est relié aussi bien au nœud qui le comporte ainsi qu'à tous les ascendants de ce nœud.

Π (t_i/e_j)	t_1	t_2	t_3	t_4	t_5
Titre	1	1	0	0	0
¬Titre	0	0	1	1	1
Résumé	1/2	1	1	1/4	0
¬Résumé	0.50	0	0	0.88	1
Chapitre	0	0	0	0.70	0.50
¬Chapitre	1	1	1	0.10	0.20
Titrechapitre	0	0	0	1	0
¬Titrechapitre	1	1	1	0	1
Paragraphe	0	0	0	0.88	1
¬Paragraphe	1	1	1	0.05	0

Tableau 3. *Distribution de possibilité Π (t_i/e_j)*

La matrice contenant les valeurs des arcs racine- nœud balise du réseau possibiliste du document 'ouvrage' est donnée dans le tableau 4 (nous prenons α = 0,6 et dl=100).

Π (ej/di)	e_1	$\neg e_1$	e_2	$\neg e_2$	e_3	$\neg e_3$	e_4	$\neg e_4$	e_5	$\neg e_5$
ouvrage	1	1	1	1	1	1	0,6	1	0,6	1
¬ouvrage	0,02	1	0,1	1	1	1	0,01	1	1	1

Tableau 4. *Distribution de possibilité Π (ej/di)*

***Remarque** : certaines valeurs considérées dans les tableaux sont prises à titre d'exemple. Elles ne correspondent pas toujours aux résultats des formules considérées. Car nous ne disposons pas de tous les paramètres pour effectuer le calcul.*

Lorsque la requête est posée, un processus de propagation est déclenché à travers le réseau modifiant les valeurs de possibilités a priori. Dans ce modèle la formule de propagation utilisée est la formule [2].

Soit une requête Q_1 composée du mot clé 'Information', Q1 = {Information}. D'après l'hypothèse 1, Π (d) = Π¬ d) = 1, quelque soit i.

Etant donnée la requête Q_1, le traitement de la formule de propagation [2] donne, on ne considère que les configurations de E qui comporte le terme de la requête 'information', en l'occurrence seules les balises e1=Titre, e2=Résumé seront considérées. Les configurations qu'il faut donc considérées sont : $\{e_1 \wedge e_2, \neg e_1 \wedge e_2, e_1 \wedge \neg e_2, \neg e_1 \wedge \neg e_2\}$. On calcule alors :

Pour di = ouvrage:

$a1=\Pi(information/e_1)*\Pi(information/e2)*\Pi(e_1/ouvrage)*\Pi(e_2/ouvrage)= 1*1*1*1=1$

$a2=\Pi(information/e_1)*\Pi(information/\neg e_2)*\Pi(e_1/ouvrage)*\Pi(\neg e_2/ouvrage)=1*0*1*1=0$

$a3=\Pi(information/\neg e_1)*\Pi(information/e_2)*\Pi(\neg e_1/ouvrage)*\Pi(e_2/ouvrage)= 0*1*1*1=0$

$a4=\Pi(information/\neg e_1)*\Pi(information/\neg e_2)*\Pi(\neg e_1/ouvrage)*\Pi(\neg e_2/ouvrage)= 0*0*1*1=0$

$\Pi(ouvrage/ Q_1).= \max(a1, a2, a3, a4)=1=a1$

Pour ¬ di = ¬ ouvrage :

$a5=\Pi(information/e_1)*\Pi(information/e_2)*\Pi(e_1/\neg ouvrage)*\Pi(e_2/\neg ouvrage)= 1*1*0.02*0.1=0.002$

$a6=\Pi(information/e_1)*\Pi(information/\neg e_2)*\Pi(e_1/\neg ouvrage)*\Pi(\neg e_2/\neg ouvrage)=1*0*0.02*1=0$

$a7=\Pi(information/\neg e_1)*\Pi(information/e_2)*\Pi(\neg e_1/\neg ouvrage)*\Pi(e_2/\neg ouvrage)=0*1*1*0.1=0$

$a8=\Pi(information/\neg e_1)*\Pi(information/\neg e_2)*\Pi(\neg e_1/\neg ouvrage)*\Pi(\neg e_2/\neg ouvrage)= 0*0*1*1=0$

$\Pi(\neg ouvrage/ Q_1).= \max(a5, a6, a7, a8)=0.002=a5$

- Pour calculer la nécessité $N (ouvrage/Q_1) = 1- \Pi(\neg ouvrage/ Q_1) = 1-0.002=0.98$
- Pour calculer la nécessité $N (\neg ouvrage/Q_1) = 1- \Pi(ouvrage/ Q_1) = 1- 1= 0$

Les documents préférés sont ceux qui ont une valeur N(di/Q) élevée parmi ceux qui ont une valeur Π (d_i/Q) élevée aussi. Si N(d_i/Q) = 0, les documents restitués sont (sans garantie d'adéquation totale) ceux qui ont une valeur Π(d_i/Q) élevée. Par conséquent, pour la requête Q_1 = {Information}, c'est la hiérarchie ouvrage-titre-résumé qui sera retournée à l'utilisateur comme réponse à sa requête.

6 Conclusion

Cet article présente un nouveau modèle théorique pour la recherche de documents XML basé sur la théorie des possibilités. L'approche proposée fournit un nouveau cadre formel pour représenter l'incertitude sur le contenu et la structure du document XML en utilisant les mesures de possibilité et de nécessité. Les mesures de possibilité et de nécessité sont utilisées pour quantifier les relations de dépendance (ou indépendance) entre les termes et les éléments du document qu'ils indexent et entre les éléments entre eux ; et permettent de restituer les documents (ainsi que les éléments) nécessairement ou possiblement pertinents étant donné une requête.

Les travaux futurs concernent l'évaluation de notre approche sur un jeu de données. Comme perspective, nous prévoyons de modéliser le degré de possibilité de la requête, étant donné les termes d'indexation, qui dépend de l'interprétation de cette dernière.

Bibliographie

[Abiteboul 97], S. Abiteboul, D. Quass, J. McHugh, J. Widom, and J.-L. Wiener. The Lorel query language for semi-structured data. *International Journal on Digital Libraries*, 1(1) :68–88, 1997.

[Abolhassani 04] M. Abolhassani and N. Fuhr. Applying the divergence from randomness approach for content-only search in xml documents. In Proceedings of ECIR 2004, Sunderland, pages 409–419, 2004.

[Baeza 99] R. Baeza-Yates., B. Ribeiro-Neto Modern Information Retrieval, Addison Wesley, 1999.

[Ben 02] N. Ben Amor, Qualitative Possibilistic Graphical Models : From Independence to Propagation Algorithms, Thèse pour l'obtention du titre de Docteur en Gestion, université de Tunis, 2002.

[Benferhat 99] S. Benferhat, D. Dubois, L. Garcia, H. Prade : Possibilistic logic bases and possibilistic graphs. In Proc. of the 15th Conference on Uncertainty in Artificial Intelligence, 57-64, 1999.

[Borgelt 00] C. Borgelt, J. Gebhardt, and R. Kruse, Possibilistic graphical models » Computational Intelligence in Data Mining, CISM Courses and Lectures 408, Springer, Wien, 51-68, 2000.

[Brini 05] A. Brini, M. Boughanem, D. Dubois. A Model for Information Retrieval Based on Possibilistic Networks. Dans : String Processing and Information Retrieval (SPIRE 2005), Buenos Aires, 2005. LNCS, Springer Verlag, p. 271-282.

[Buneman 96], P. Buneman, S. Davidson, G. Hillebrand, D. Suciu. A query language and optimization techniques for unstructured data. In *Proceedings of ACM SIGMOD International Conference on Management of Data, Montréal*, pages 505–516, 1996.

[Ceri 99] S. Ceri, S. Comai, E. Damiani, P. Fraternali, S. Paraboschi, and L. Tanca. XML-GL : A graphical language for querying and restructuring WWW data. In *Proceedings Of the 8th Int. WWW Conference, WWW8, Toronto, Canada*, May 1999.

[Chamberlin 00], D. Chamberlin, J. Robie, D. Florescu. Quilt : An XML query language for heterogeneous data sources. In *Proceedings of the 3rd Internation Workshop on World Wide Web and databases, Dallas, USA*, pages 1–25, 2000.

[Govert 02] N. Govert, M. Abolhassani, N. Fuhr, and K. Grossjohann. Contentoriented XML retrieval with hyrex. In Proceedings of the first INEX Workshop, Dagstuhl, Germany, 2002.

[Hayashi 00] Y. Hayashi, J. Tomita, G. Kikoi, Searching text-rich XML documents with relevance ranking. In Proc ACM SIGIR 2000 Workshop on XML and IR (pp. 27-35). Athens 2000.

[Kamps 03] Kamps J., Marx M., De Rijke M., Sigurbjörnsson B., XML Retrieval : What to retrieve ? ACM SIGIR Conference on Research and Development in Information Retrieval, p.409-410, 2003.

[Lalmas 97] M. Lalmas. Dempster-Shafer's Theory of Evidence Applied to Structured Documents: Modelling Uncertainty. In Proceedings of the 20th Annual International ACM SIGIR, pages 110–118, Philadelphia, PA, USA. ACM. (1997)

[Lalmas 04] M. Lalmas, P. Vannoorenberghe. Indexation et recherche de documents XML par les fonctions de croyance, Première Conférence en Recherche d'Information et Applications, CORIA'2004, pp 143-160 (2004).

[Levy 98], A. Levy, M. Fernandez, D. Suciu, D. Florescu, and A. Deutsch. XMLQL: A query language for XML. Technical report, World Wide Web Consortium technical report, Number NOTE-xml-ql-19980819, 1998.

[Ogilvie 03] P. Ogilvie and J. Callan. Using language models for flat text queries in xml retrieval. In Proceedings of INEX 2003 Workshop, Dagstuhl, Germany, pages 12–18, December 2003.

[Piwowarski 02] B. Piwowarski, G.E. Faure, P. Gallinari, « Bayesian Networks and INEX », In INEX 2002 Workshop Proceedings, p. 149-153, Germany, 2002.

[Rölleke 02] T. Rölleke. M. Lalmas., G. Kazai., I. Ruthven, S. Quicker. The accessibility Dimension for Structured Document Retreival', BCS-IRSG European Conference on Information Retrieval (ECIR), Glasgow, Mars 2002.

[Salton 83] G. Salton and M. McGill. *Introduction to modern information retrieval.*McGraw-Hill Int. Book Co, 1983.

[Sauvagnat 05] K. Sauvagnat., Modèle flexible pour la Recherche d'Information dans des corpus de documents semi-structurés, thèse de Doctorat de l'Université Paul Sabatier, Juillet 2005.

[Sigurbjornsson 03] B. Sigurbjornsson, J. Kamps, and M. de Rijke. An element-based approach to xml retrieval. In Proceedings of INEX 2003 workshop, Dagstuhl, Germany, december 2003.

[Schlieder 02] T. Schlieder, H. Meuss "Querying and ranking XML documents". *Journal of the American Society for Information Science and Technology*, 53(6) : 489-503, 2002.

[Zadeh 78] L. A. Zadeh, Fuzzy Sets as a Basis for a theory of possibility, In Fuzzy Sets and Systems, 1 :3-28, 1978.