
Indexation et recherche de documents XML par les fonctions de croyance

Mounia Lalmas* et Patrick Vannoorenberghe**

* *Department of Computer Science
Queen Mary University of London
London E1 4NS, United Kingdom
mounia@dcs.qmul.ac.uk*

** *Perception Systèmes Information, FRE 2645 CNRS
Université de Rouen, Faculté des Sciences
Place Emile Blondel, 76821 Mont Saint Aignan, France
Patrick.Vannoorenberghe@univ-rouen.fr*

RÉSUMÉ. Dans cet article, nous nous intéressons à la recherche de documents XML. Un cadre générique qui permet la représentation de connaissances partielles dans les processus d'indexation et de recherche est tout d'abord présenté. Ce modèle est basé sur la théorie des fonctions de croyance et permet de décrire plusieurs formes d'incertitude sur le contenu et la structure des documents XML. Par ce biais, la méthodologie autorise l'utilisation de requêtes qui permettent la spécification de contraintes sur la structure des documents recherchés. Un exemple d'une telle stratégie de recherche est proposé de façon à illustrer la méthodologie présentée.

ABSTRACT. This paper focuses on XML document retrieval. A generic framework which allows us to represent uncertain knowledge in the indexing and retrieval processes is first presented. This model is based on belief functions theory and can describe various forms of uncertain knowledge about the content and the structure of XML documents. Retrieval is performed within this framework and can include the use of queries that allow the specification of structural conditions. An example of such strategy is then presented in order to illustrate our approach.

MOTS-CLÉS : Recherche d'Information, Indexation, Document XML, Fonctions de croyance.

KEYWORDS: Information Retrieval, Indexation, XML document, Belief functions.

1. Introduction

Avec l'adoption de XML comme format standard pour les documents structurés, la recherche d'information requiert des systèmes compatibles à la gestion de requêtes basées sur le contenu mais aussi sur la structure des documents XML [BLA 03a, FUH 03]. Au lieu d'évaluer la pertinence d'un document dans sa globalité, de tels systèmes de fouille doivent permettre la recherche des composants spécifiques du document. Autrement dit, on s'attache à rechercher des composants pertinents à différents niveaux de granularité. Cette approche se démarque des méthodes qui consistent à rechercher les éléments de granularité fixé [WIL 94]. De plus, ils doivent être capables d'appréhender l'incertitude intrinsèque associée au processus de recherche d'information pour permettre un accès pertinent aux documents XML. Quantifier l'incertitude dans ce contexte ne se situe pas qu'au niveau du contenu du document mais aussi de sa structure. En ce qui concerne l'incertitude liée au contenu, il s'agit de quantifier la façon dont un terme peut décrire le contenu d'un élément XML. Au niveau de la structure, on cherche à évaluer la capacité que possède un élément XML à fournir une information pertinente à l'utilisateur.

Dans cette optique, un certain nombre de méthodes de recherche de documents XML ont été proposées [BER 02, BOA 02]. Ces techniques fournissent un accès aux documents XML en se basant sur la mise en correspondance de chaînes de caractères ou d'expressions régulières qui ne suffisent généralement pas aux recherches d'information sur des bases de documents XML. Des exemples de telles approches peuvent être trouvées dans [FUH 01, THE 02]. La recherche orientée 'pertinence' peut être menée en utilisant des techniques de recherche d'information comme par exemple la pondération des termes [BAE 99] où seule l'incertitude liée au contenu du document y est considérée. Récemment, plusieurs approches considérant l'incertitude quant à la structure ont été proposées [BLA 03b]. Notre travail a pour but d'uniformiser dans un même cadre théorique la gestion des incertitudes liées au contenu et à la structure.

Dans cet article, nous décrivons un cadre générique qui permet de représenter l'incertitude au niveau du contenu et de la structure de documents XML. On suppose ici que la collection des documents traités est homogène (c'est-à-dire conformes à une même DTD). Ce cadre est basé sur la théorie des fonctions de croyance et permet de modéliser plusieurs formes d'incertitude dans un contexte de recherche de documents XML. L'approche proposée est la suite de travaux présentés dans [LAL 97], qui utilisaient également les fonctions de croyance pour construire un modèle de documents structurés. Cependant, dans l'approche précédente seule l'incertitude au niveau du contenu était modélisée. La section 2 donne un aperçu des concepts mathématiques relatifs aux fonctions de croyance. Un exemple de document XML est présenté à la section 3 et nous permet d'introduire les différentes notations utilisées pour décrire les aspects de la recherche et de l'indexation qui sont formellement modélisés dans l'approche proposée. Nous présentons le cadre théorique utilisé pour représenter l'incertitude relative au contenu et à la structure des documents XML dans la section 4. La stratégie de recherche d'information est présentée à la section 5 et illustrée par un exemple à la section 6.

2. Théorie des fonctions de croyance

Soit Ω un ensemble fini de valeurs possibles pour une variable x . Cet ensemble est généralement appelé cadre de discernement et correspond au recensement exhaustif des valeurs potentielles prises par la variable d'intérêt. Pour représenter la connaissance partielle sur la valeur prise par x , un cadre mathématique qui permet de représenter l'incertitude sur la source d'information est requis. Un modèle général peut être défini en associant à chaque sous-ensemble A de Ω une mesure qui représente la confiance associée au sous-ensemble A étant donné l'état de connaissance. Dans ce modèle, A correspond à l'ensemble des valeurs possibles pour la variable x . L'incertitude est capturée par la mesure de confiance allouée à A . À partir de considérations mathématiques posées par Dempster, Shafer [SHA 76] a montré la capacité qu'offrent les fonctions de croyance pour représenter la connaissance incertaine. L'utilité des fonctions de croyance, comme une alternative aux probabilités subjectives, a été démontrée de manière axiomatique par Smets [SME 94] avec le *Modèle des Croyances Transférables* (MCT) donnant ainsi une interprétation claire et cohérente aux concepts sous-jacents à la théorie.

2.1. Connaissance incertaine et fonctions de croyance

La connaissance partielle sur la valeur prise par la variable x peut être représentée par une **fonction de masse** (basic belief assignment : bba) m qui est définie comme une fonction de 2^Ω dans $[0, 1]$ qui vérifie $\sum_{A \subseteq \Omega} m(A) = 1$. Les sous-ensembles $A \subseteq \Omega$ tels que $m(A) > 0$ sont les éléments focaux de m . Chaque élément focal A est un ensemble de valeurs possibles pour x , et la quantité $m(A)$ peut être interprétée comme la part de croyance allouée à A étant donné l'état de la connaissance. L'ignorance complète peut être codée par la fonction de masse vide $m(\Omega) = 1$. Une fonction de masse m est dite normale si $m(\emptyset) = 0$. Si tous les éléments focaux sont des singletons alors la fonction de masse est dite **Bayésienne**. Si les éléments focaux sont emboîtés (c'est-à-dire linéairement ordonnés par inclusion), la fonction de masse est appelée **consonante**. Une fonction de masse m peut être représentée de façon équivalente par une mesure floue non-additive de $2^\Omega \rightarrow [0, 1]$ appelée fonction de croyance, notée bel et définie par :

$$bel(A) \triangleq \sum_{\emptyset \neq B \subseteq A} m(B) \quad \forall A \subseteq \Omega. \quad (1)$$

La quantité $bel(A)$ peut être interprétée comme la croyance totale pour que l'hypothèse A soit vérifiée. Il est à remarquer que les fonctions m et bel sont équivalentes [SHA 76] et qu'elles peuvent être vues comme deux facettes de la même information.

Lorsque la source d'information dont la fonction de croyance est extraite n'est pas totalement fiable, il est possible d'introduire une opération d'affaiblissement. Ainsi, une fonction de masse affaiblie, notée m_α , peut se déduire de m par :

$$m_\alpha(A) = \alpha m(A) \quad \forall A \subseteq \Omega, A \neq \Omega \quad (2)$$

$$m_\alpha(\Omega) = 1 - \alpha + \alpha m(\Omega) \quad (3)$$

avec $0 \leq \alpha \leq 1$. Dans ce contexte, le coefficient α , qui représente une sorte de méta-connaissance concernant la fiabilité de la source, permet de transférer une partie de la croyance vers l'ensemble Ω .

2.2. Règle de combinaison de Dempster

Supposons que nous disposons de deux allocations de masse m_1 et m_2 définies sur le même référentiel Ω représentant deux sources d'information sur la valeur de la variable \mathbf{x} . Ces deux fonctions peuvent être agrégées par un opérateur de combinaison conjonctif noté \odot . Le résultat de cette opération conduit à une fonction de croyance unique à laquelle correspond une fonction de masse, notée m_{\odot} , qui peut être définie par :

$$m_{\odot}(A) = (m_1 \odot m_2)(A) \triangleq \sum_{B \cap C = A} m_1(B) m_2(C) \quad \forall A \subseteq \Omega. \quad (4)$$

Cette règle conjonctive est parfois dénommée règle de combinaison de Dempster non normalisée. Si nécessaire, l'hypothèse de normalisation $m_{\odot}(\emptyset) = 0$ peut être retrouvée en divisant chaque masse par un coefficient adéquat. L'opérateur résultant, qui est connu sous le nom de règle de Dempster et noté m_{\oplus} , est défini par :

$$m_{\oplus}(A) = (m_1 \oplus m_2)(A) \triangleq \frac{(m_1 \odot m_2)(A)}{1 - m_{\odot}(\emptyset)} \quad \forall A \subseteq \Omega \quad (5)$$

où la quantité $m_{\odot}(\emptyset)$ est le degré de conflit entre les fonctions m_1 et m_2 . L'utilisation de la règle de Dempster est possible si les fonctions de masse m_1 et m_2 ne sont pas en conflit total c'est-à-dire s'il existe deux éléments focaux B et C respectivement de m_1 et m_2 tels que $B \cap C \neq \emptyset$. Cette règle possède des propriétés intéressantes (associativité, commutativité, non-idempotence) et son utilisation a été démontrée théoriquement par plusieurs auteurs [SME 94].

2.3. Principes de marginalisation et d'extension

Soit une fonction de masse m^Ω définie sur le produit Cartésien $\Omega = \Omega_1 \times \Omega_2$. La fonction de masse marginale $m^{\Omega \downarrow \Omega_1}$ sur Ω_1 est définie, pour tout $A \subseteq \Omega_1$, par :

$$m^{\Omega \downarrow \Omega_1}(A) \triangleq \sum_{\{B \subseteq \Omega \mid Proj(B \downarrow \Omega_1) = A\}} m^\Omega(B) \quad (6)$$

où $Proj(B \downarrow \Omega_1)$ représente la projection de B sur Ω_1 , qui est définie par :

$$Proj(B \downarrow \Omega_1) \triangleq \{\omega_1 \in \Omega_1 \mid \exists \omega_2 \in \Omega_2; (\omega_1, \omega_2) \in B\}. \quad (7)$$

Le concept de marginalisation permet de passer à un cadre de discernement de granularité plus fine. Le concept d'extension permet quant à lui d'élargir le cadre de discernement. Etant donnée une fonction de masse m^{Ω_1} sur Ω_1 , son extension sur $\Omega = \Omega_1 \times \Omega_2$ est définie pour tout $B \subseteq \Omega$ par

$$m^{\Omega_1 \uparrow \Omega}(B) \triangleq \begin{cases} m^{\Omega_1}(A), & \text{si } B = A \times \Omega_2 \text{ pour } A \subseteq \Omega_1 \\ 0 & \text{sinon.} \end{cases}$$

Cette définition de l'extension provient du principe d'engagement minimal qui formalise l'idée qu'il ne faut pas accorder plus de confiance à une proposition que celle qui lui est justifiée. La combinaison conjonctive sur $\Omega = \Omega_1 \times \Omega_2$ de deux fonctions de masse m^{Ω_1} et m^{Ω_2} peut être obtenue en combinant leurs extensions respectives sur Ω . On obtient

$$(m^{\Omega_1} \odot m^{\Omega_2})(A \times B) = m^{\Omega_1}(A)m^{\Omega_2}(B) \quad (8)$$

pour tous les sous-ensembles $A \subseteq \Omega_1$ et $B \subseteq \Omega_2$.

2.4. Fonction de croyance et sous-ensemble flou

Soit F un sous-ensemble flou associé à un référentiel ω et π la distribution de possibilité correspondante. Ces deux entités sont reliées par l'équation suivante :

$$\pi(\omega) = F(\omega) \quad \forall \omega \in \Omega$$

où $F(\omega)$ est le degré d'appartenance de ω à F . Chaque nombre $\pi(\omega)$ est interprété comme le degré de possibilité que la variable d'intérêt x soit égale à ω , à partir de la proposition floue ' x est F '. La mesure de possibilité associée Π est alors obtenue pour tout sous-ensemble discret A de Ω par :

$$\Pi(A) \triangleq \max_{\omega \in A} \pi(\omega),$$

tandis que la mesure de nécessité correspondante N est donnée par :

$$N(A) = \Pi(\Omega) - \Pi(\overline{A}).$$

N est formellement identique à une fonction de croyance consonante, c'est-à-dire une fonction de croyance dont les éléments focaux sont emboîtés. Les éléments focaux de N sont les α -coupes de F et la fonction de masse correspondante est définie comme suit. Soit $\pi_1 > \dots > \pi_r$ les valeurs distinctes prises par π classées dans l'ordre décroissant avec $\pi_{r+1} = 0$ par convention. Soit A_i la π_i -coupe de F . Pour tout sous-ensemble A de Ω non vide, nous avons :

$$m_F(A) \triangleq \begin{cases} \pi_i - \pi_{i+1} & \text{si } A = A_i, i = 1, \dots, r \\ 0 & \text{sinon.} \end{cases}$$

2.5. Transformation pignistique

Soit m^Ω une fonction de masse définie sur Ω . Smets [SME 94] propose de transformer m^Ω en une fonction de probabilité $BetP_{m^\Omega}$ sur Ω (appelée fonction de probabilité pignistique) définie pour tout $\omega \in \Omega$ par :

$$BetP_{m^\Omega}(\omega) = \sum_{X \subseteq \Omega | \omega \in X} \frac{m^\Omega(X)}{|X|}, \quad (9)$$

où $|X|$ est la cardinalité de $X \subseteq \Omega$. Dans cette transformation, la masse de croyance $m^\Omega(X)$ est uniformément distribuée parmi les éléments de X . Il est facile de montrer que $BetP_{m^\Omega}$ est une fonction de probabilité. D'autres fonctions peuvent être obtenues à partir de $BetP_{m^\Omega}$ pour tout sous-ensemble $A \subseteq \Omega$ en utilisant :

$$BetP_{m^\Omega}(A) = \sum_{\omega \in A} BetP_{m^\Omega}(\omega), \quad \forall A \subseteq \Omega. \quad (10)$$

Ces fonctions pignistiques peuvent être utilisées pour la prise de décision en environnement incertain.

3. Incertitude inhérente aux documents XML

Dans le format XML, le texte est encapsulé entre des balises de début et de fin dont le nom fournit une indication sur le type d'information qu'elles renferment. Le texte et ses balises correspondent aux éléments XML. Les éléments, qui sont généralement emboîtés, peuvent être assignés à des attributs qui sont donnés dans la balise de début. Un exemple de document XML relatif à un Compact-Disc (CD) nous permet d'illustrer nos propos.

```
<CD serial="0602498097335" disc-length="58 :13">
<company>AM Records</company>
<artist>Sting</artist>
<title>Sacred love</title>
<genre>Pop</genre>
<date>2003</date>
<song>
<title>Inside</title>
<length>4 :47</length>
<chorus>Outside the stars are turning...</chorus>
<verse>Inside the doors are scaled to...</verse>
<verse>Inside my head's a box of stars...</verse>
</song>
<song>
<title>Send your love</title>
<length>4 :38</length>
<chorus>Send your love into the future...</chorus>
```

```
<verse>Finding the world in the smallness...</verse>
<verse>Inside your mind is a relay...</verse>
</song>
```

</CD> De plus, une DTD (Document Type Definition) doit être spécifiée en ce qui concerne la syntaxe d'une collection de documents XML. Ainsi, un document XML est dit 'valide' s'il est conforme à la DTD correspondante. Des expressions XPath sont utilisées pour accéder aux éléments XML qui composent le document. Par exemple, pour se référer au titre de la première chanson d'un Compact-Disc (CD), l'expression XPath suivante est utilisée : `CD/song[1]/title`. Pour accéder à tous les titres du CD (c'est-à-dire le titre du CD et les titres des différentes chansons le composant) l'expression : `CD//title` est employée. Pour avoir accès à tous les éléments-fils de l'élément CD, l'expression XPath `CD/*` est utilisée. Finalement, les éléments descendants de l'élément CD sont accessibles via l'expression : `CD/**`. Les types de ces quatre groupes d'éléments XML sont respectivement : `CD/song/title`; `CD/title`, `CD/song/title`; `CD/company`, `CD/artist`, ..., `CD/date`, `CD/song` et tous les types précédents, `CD/song/title`, `CD/song/length`,... Ces types peuvent être directement extraits de la DTD associée à la collection de documents.

Certaines notations et terminologies sont introduites de manière à faire référence aux différentes entités constituant un document XML. Pour chaque collection de documents XML, on suppose connu :

1) un ensemble de n termes dans la collection, après 'suppression des mots vides, radicalisation, ...' [BAE 99] noté $T = \{t_1, \dots, t_n\}$.

2) un ensemble de r types d'éléments qui peuvent être retrouvés (on suppose dans ce travail que seuls les types significatifs en terme de contenu peuvent être retrouvés) $Y = \{y_1, \dots, y_r\}$. Cet ensemble est déduit partiellement de la DTD associée au jeu de données.

3) un ensemble de s éléments $E = \{e_1, \dots, e_s\}$, qui correspondent aux éléments eux-mêmes et doivent être d'un type défini dans Y . Ceci est modélisé par la fonction $of_type : E \rightarrow Y$. Ce modèle correspond aux expressions XPath simples de la forme `element[.]/element[.]/element[.]/...`, chacun accédant à un élément.

4) les relations parent-fils entre les éléments qui sont formalisées par la relation $part_of : E \rightarrow \wp(E)$, tel que pour un élément $e \in E$, $|part_of(e)| \leq 1$. Deux cas sont à considérer. Le premier pour lequel $part_of(e) = \emptyset$ signifie que l'élément e est un élément racine qui n'a pas de parent. Le second, $part_of(e) = \{e'\}$ exprime le fait que l'élément e' est le parent de l'élément e . Comme dans la majorité des cas pour les documents XML, un élément possède au moins un parent (du fait de la structure hiérarchique). Ces relations sont extraites des expressions XPath utilisées pour accéder aux différents éléments constituant le document. Par exemple, pour l'élément `CD/song[1]/verse[2]`, on obtient $part_of(CD/song[1]/verse[2]) = \{CD/song[1]\}$ puisque `CD/song[1]` est l'élément parent de `CD/song[1]/verse[2]`.

5) en utilisant la relation $part_of$ définie précédemment, on peut introduire la relation qui matérialise les fils d'un élément comme une fonction $children : E \rightarrow$

$\wp(E)$, telle que pour un élément $e \in E$, $children(e) = \{e' \in E \mid part_of(e') = e\}$. Dans notre exemple, on obtient $CD/song[1]/verse[2] \in CD/song[1]$.

Ces notations sont relatives aux éléments d'un document XML, leurs types et leurs structures, comme par exemple les relations parents-fils. Nous introduisons ensuite la terminologie relative à l'incertitude concernant la représentation du contenu et de la structure des éléments XML. Dans le domaine de la recherche d'informations, les termes utilisés pour représenter le contenu d'un document sont pondérés de manière à capturer la façon dont ils décrivent le contenu [BAE 99]. Le même principe peut être appliqué à la recherche de documents XML. Les poids sont généralement calculés en utilisant l'information de fréquence d'apparition des termes (le nombre de fois qu'un terme apparaît dans un élément) et la fréquence inverse du document (le nombre d'éléments dans lesquels le terme apparaît). Pour représenter ces deux types de fréquence, nous définissons les deux fonctions suivantes :

- 1) pour chaque élément $e \in E$ et chaque terme $t \in T$, la fonction $tf : E \times T \rightarrow [0, 1]$ donne la fréquence d'apparition du terme comme un réel dans $[0, 1]$.
- 2) pour chaque terme $t \in T$, la fonction $ief : T \rightarrow [0, 1]$ donne la fréquence inverse d'apparition d'un terme t dans la collection comme un réel dans $[0, 1]$.

En ce qui concerne la structure des documents XML, l'incertitude peut être interprétée de façon similaire à l'incertitude sur les termes ; c'est-à-dire qu'il s'agit de quantifier la capacité d'un type d'élément à fournir une information pertinente à l'utilisateur. En effet, les types ne doivent pas être traités de façon uniforme car certains peuvent s'avérer plus importants que d'autres en terme d'indexation. Par exemple, l'élément SECTION dans un article peut fournir une information plus riche en terme de performance d'indexation que l'élément TITRE ; même chose en ce qui concerne l'élément PARAGRAPHE contenu dans un RESUME par rapport à un PARAGRAPHE dans un élément SECTION. De la même façon, des éléments plus larges peuvent amener une meilleure indexation que des éléments dont le champ serait plus réduit [KAM 03]. Nous supposons qu'une telle information est disponible pour tous les types et qu'elle peut être modélisée au travers de la fonction $typew : Y \rightarrow [0, 1]$ qui modélise l'impact associé au type $y \in Y$ sous la forme d'un réel dans $[0, 1]$.

4. Représentation de documents XML

La contribution essentielle de ce travail concerne la représentation du contenu et de la structure de documents XML dans un même formalisme. Nous avons choisi d'utiliser la théorie des fonctions de croyance pour sa capacité à représenter des connaissances partielles (de l'ignorance totale à la connaissance complète). On suppose tout d'abord qu'il est possible de quantifier l'incertitude sur le contenu d'un élément XML par l'intermédiaire d'une fonction de masse définie sur l'ensemble des termes T . Cette allocation de masse sera détaillée dans la section 4.1. L'incertitude concernant la structure du document peut être formalisée quant à elle par une fonction de masse définie sur l'ensemble des types d'élément Y (voir section 4.2). De façon à obtenir une re-

présentation du document qui englobe le contenu et la structure associées à leurs incertitudes respectives, il suffit de combiner ces deux fonctions de masse sur le produit Cartésien $\Omega = T \times Y$ en utilisant l'équation (8). Dans les sections suivantes, nous nous attardons sur la représentation des incertitudes liées au contenu et à la structure d'un document XML.

4.1. Représentation du contenu

L'ensemble des termes T obtenu après l'application des techniques classiques d'indexation (suppression des mots vides, radicalisation [BAE 99]) est choisi comme cadre de discernement pour représenter le contenu. Ayant observé l'élément $e \in E$, une fonction de masse notée $m^T[e]$ est définie sur T pour chaque terme $t \in T$ tel que $tf(e, t) > 0$, $m^T[e](\{t\}) \neq 0$. Ce choix s'explique par le fait qu'un terme qui apparaît dans un élément constitue par nature un élément focal de $m^T[e]$.

Pour chaque terme $t \in T$ tel que $m^T[e](\{t\}) \neq 0$, $m^T[e](\{t\})$ s'obtient à partir de $tf(e, t) > 0$ et/ou $ief(t)$. L'incertitude totale, c'est-à-dire la masse $m^T[e](T)$ allouée à T , représente la connaissance disponible sur la fiabilité du processus d'indexation. La façon dont cette allocation $m^T[e](\cdot)$ est calculée permet de satisfaire l'hypothèse $\sum_{A \subseteq T} m^T[e](A) = 1$ de manière à ce que $m^T[e]$ soit une fonction de masse.

Dans ce contexte, chaque fonction $m^T[e]$ permet d'obtenir une représentation de chaque élément XML en terme de contenu. En recherche d'informations, il a été montré [RÖL 02] que les performances du système pouvaient être améliorées de façon significative si l'on représentait un élément par son propre contenu mais aussi par celui de ces éléments-fils. Plus précisément, en recherche de documents structurés, le contenu d'un élément-parent est souvent caractérisé comme la combinaison de son contenu propre et de celui de ses éléments-fils. Ceci peut être rendu possible dans notre cadre en utilisant la règle de combinaison de Dempster comme elle a déjà été appliquée dans [LAL 97]. Dans cet article, la contribution provient du fait que les éléments-fils n'ont pas tous le même impact sur le contenu d'un élément-parent. La contribution de chaque élément-fils doit donc être formalisée. Par exemple, dans un ensemble d'ARTICLES, un RESUME permet d'avoir une vision plus globale du contenu de l'article qu'une CONCLUSION. Il paraît donc clair qu'il existe une différence entre le poids associé à un type d'élément (son importance dans le processus de recherche) puisque nous parlons de l'importance relative d'un élément envers ses éléments-parents en terme de représentation agrégée.

Pour faciliter la compréhension, on considère un élément e qui possède deux éléments fils e_1 et e_2 , c'est-à-dire $children(e) = \{e_1, e_2\}$. A chacun de ces éléments correspond un corpus d'évidence et sa fonction de masse respective $m^T[e]$, $m^T[e_1]$ et $m^T[e_2]$. On suppose donné l'impact du contenu de chaque élément-fils e_1 et e_2 sur la

représentation agrégée du contenu de e . Ceci peut être modélisé par des coefficients d'affaiblissement α_{e_i} pour $i = 1, 2$ (voir équations (2) et (3)) :

$$\check{m}^T[e_i](A) = \alpha_{e_i} m^T[e_i](A) \quad \forall A \subseteq T, A \neq T \quad (11)$$

$$\check{m}^T[e_i](T) = 1 - \alpha_{e_i} + \alpha_{e_i} m^T[e_i](T). \quad (12)$$

Dans ces équations, chaque paramètre α_{e_i} représente une sorte de méta-connaissance sur la contribution relative d'un élément-fils e_i . Ces paramètres peuvent être ajustés à partir d'études sur les utilisateurs (élicitation directe), expérimentalement ou par des méthodes d'apprentissage. La détermination de ces différentes valeurs sort du cadre de cet article et on supposera dans la suite que ces coefficients sont donnés. Les fonctions de croyance affaiblies obtenues à partir des équations (11) et (12) peuvent être combinées avec la règle de Dempster de façon à obtenir une représentation d'un élément basé sur ses éléments-fils de contribution respective. Ceci conduit à :

$$m^T[\text{children}(e)] = \bigoplus_{e' \in E | \text{part_of}(e')=e} \check{m}^T[e']. \quad (13)$$

Ensuite, il s'agit de combiner le résultat avec la fonction de masse $m^T[e]$ relative à l'élément e lui-même :

$$m^T[e, \text{children}(e)] = m^T[\text{children}(e)] \oplus m^T[e]. \quad (14)$$

Ce modèle permet d'encapsuler l'incertitude totale existant dans un élément XML en terme de contenu. Puisqu'il est basé également sur les contributions respectives des éléments-fils, il permet aussi de prendre en considération la structure d'une certaine manière.

4.2. Représentation de la structure

Pour représenter la structure d'un document, l'idée consiste à définir un corpus d'évidence qui permet de quantifier l'importance de chaque type d'élément $y \in Y$ dans la collection. Pour ce faire, le cadre de discernement Y est défini comme l'ensemble des types d'éléments. La fonction de masse m_1^Y permet de capturer l'importance de chaque type d'élément. Cette fonction peut être déduite de la fonction $typew : Y \rightarrow [0, 1]$ en supposant qu'une normalisation spécifique est appliquée aux valeurs $typew(\cdot)$ tel que $\max_y typew(y) = 1$. Cette normalisation permet ainsi d'obtenir une distribution de possibilité. Néanmoins, ces valeurs peuvent être obtenues par élicitation directe à partir d'avis d'experts (en leur posant la question : Quel est le degré de possibilité concernant l'importance des éléments de type y ?) ou peuvent être calculées à partir d'opinions d'experts combinées. Une distribution de possibilité est formellement équivalente à une fonction de croyance consonante c'est-à-dire une fonction de croyance dont les éléments focaux sont emboîtés.

Ainsi, à partir de la distribution obtenue après normalisation, nous obtenons une fonction de masse m_1^Y dont un exemple est illustré au tableau 1 pour un ensemble

.	$typew(\cdot)$	$m_1^Y(\cdot)$	$typew(\cdot)$	$m_1^Y(\cdot)$
y_1	0.3	0.0	0.3	0.0
y_2	1.0	0.7	1.0	0.7
y_3	0.0	0.0	0.1	0.0
y_1, y_2	nd	0.3	nd	0.2
y_1, y_3	nd	0.0	nd	0.0
y_2, y_3	nd	0.0	nd	0.0
Y	nd	0.0	nd	0.1

Tableau 1. Fonctions de masse m_1^Y définies sur Y calculées à partir de la fonction $typew$ représentant l'importance de chaque élément au sein de la structure.

Y composé de trois types d'éléments. La fonction de masse m_1^Y quantifie l'importance de chaque élément de type $y \in Y$. Comme exemple, supposons que l'on possède un document XML relatif à un article scientifique avec les éléments de type $y_1 = \text{ARTICLE/ABSTRACT/PARA}$ et $y_2 = \text{ARTICLE/SECTION/PARA}$. Dans ce cas, le degré de croyance $m_1^Y(y_1, y_2)$ permet de quantifier l'importance d'un élément de type PARA inclus dans un élément de type ABSTRACT ou SECTION.

A partir de cette première modélisation, nous proposons d'étendre la connaissance par un corpus d'évidence relatif à une requête qui présenterait des conditions sur la structure permettant ainsi la gestion de requêtes imposant des contraintes sur la structure. Supposons que cet état de connaissance soit modélisé par une fonction de masse, notée m_2^Y définie sur le même cadre de discernement Y utilisé pour encoder la structure. Cette fonction peut être déduite d'un système d'apprentissage basé sur l'étude d'un historique de requêtes, ajustée par les utilisateurs ou extraite de la requête courante. Cette dernière option nous semble la plus séduisante pour définir complètement la fonction de masse m_2^Y . Ainsi, une requête basée sur la recherche d'une structure simple peut être encodée en allouant une partie de la masse à l'élément de type concerné par la requête. Des structures plus complexes basées sur des intersections d'éléments de différents types peuvent également être modélisées. Cependant, si l'utilisateur ne souhaite pas prendre en compte cette option, la fonction sera définie par $m_2^Y(Y) = 1$ représentant ainsi l'ignorance totale sur la connaissance de la structure relative à la requête.

Finalement, pour obtenir une représentation qui tienne compte de la structure d'un document XML et de la requête, il nous faut combiner les fonctions m_1^Y et m_2^Y par la règle de combinaison de Dempster pour obtenir $m^Y = m_1^Y \oplus m_2^Y$. La fonction de masse m^Y correspond à l'importance des types dans le document XML mais aussi au type d'éléments cibles spécifiés dans la requête. On peut remarquer, qu'à ce niveau, le modèle ne permet de spécifier que l'élément cible de la requête. En effet, l'élément cible correspond au type d'élément qui doit être retourné par le système (recherche CHORUS par exemple). Nos travaux futurs s'orientent vers des conditions quant au contenu dans la structure comme par exemple rechercher SONG qui possèdent CHORUS

contenant 'love'. Dans cet exemple, SONG est l'élément cible tandis que la condition sur le contenu correspond à CHORUS contenant le terme 'love'.

4.3. Incertitude liée au document XML

Dans ce paragraphe, nous recensons l'information disponible qui a été extraite du contenu et de la structure de chaque document dans la collection. Ainsi, à notre disposition :

- un ensemble de fonctions de masse $m^T[e, children(e)]$ qui permettent de quantifier le contenu d'un document XML ayant observé chaque élément $e \in E$ de type $y \in Y$ (on rappelle que $y = of_type(e)$); ces corpus d'évidence¹ sont basés sur l'observation de l'élément e lui-même et de ses éléments-fils e_i de contribution respective α_{e_i} ,

- une fonction de masse m^Y représentant la structure du document.

A partir de ces informations, il est possible d'obtenir une fonction de masse m^Ω définie sur le produit Cartésien $\Omega = T \times Y$ qui peut s'exprimer par :

$$m^\Omega(A \times B) = (m^T[e] \odot m^Y)(A \times B) \quad (15)$$

$$= m^T[e](A) \cdot m^Y(B) \quad (16)$$

pour tous les sous-ensembles non vides $A \subseteq T$ et $B \subseteq Y$. Les sous-ensembles du cadre de discernement Ω sont des configurations par rapport à T et Y . La masse $m^\Omega(t \times y)$ représente le degré de croyance dans la proposition 'le terme t indexe l'élément de type y '. Cette fonction de masse étant définie sur le produit Cartésien peut être difficile à manipuler. Une autre manière d'obtenir un corpus d'évidence relatif au contenu d'un document XML ayant observé sa structure consiste à calculer la fonction de masse $m^T[e, m^Y]$. Ceci est possible en utilisant les concepts de marginalisation et d'extension pour obtenir :

$$m^T[e, m^Y] = (m^\Omega \oplus m^{Y \uparrow \Omega}) \downarrow T. \quad (17)$$

La fonction obtenue $m^T[e, m^Y]$ est alors définie sur l'ensemble des termes T (facile à interpréter) et quantifie la croyance pour que le terme $t \in T$ indexe l'élément e ayant observé son contenu, celui de ses éléments-fils ainsi que sa structure. Nous expliquons dans la section suivante comment cette fonction de masse est utilisée dans la phase de recherche.

5. Recherche de documents XML

Dans la section précédente, nous avons utilisé la théorie des fonctions de croyance pour parvenir à une représentation de documents XML où les incertitudes liées au

1. Par mesure de simplification, ces fonctions de masse seront notées $m^T[e]$ dans la suite de cet article.

contenu et à la structure ont été modélisées. Dans cette section, ce cadre formel est appliqué à la recherche de documents. La principale différence entre la recherche d'informations standard et la recherche de documents XML provient du fait que les requêtes peuvent contenir des conditions sur le contenu et la structure. Ainsi, la réponse à une requête peut se composer d'éléments XML arbitraires. Il existe deux types de requêtes pour accéder aux documents XML :

1) Requêtes basées sur le Contenu (RC) : sont des requêtes qui ignorent la structure du document et ne contiennent uniquement que des conditions relatives au contenu c'est-à-dire sur ce qu'un élément contient en terme d'indexation. Ce type de requête peut s'avérer utile pour des utilisateurs qui ne se soucient pas de la structure des éléments résultats ou tout simplement qui ne sont pas familiés avec la structure exacte des documents XML.

2) Requêtes basées sur le Contenu et la Structure² (RCS) : sont des requêtes qui contiennent des références à la structure XML en restreignant le contexte d'intérêt et/ou limitant les concepts de certaines recherches.

Les paragraphes qui suivent (5.1 et 5.2) expliquent en détail la façon dont la recherche est menée pour ces deux types de requêtes.

5.1. Requête basée sur le Contenu

Puisque les Requêtes basées sur le Contenu ignorent totalement la structure du document, il semble naturel de penser que le corpus d'évidence dont la fonction de masse $m^T[e]$ est dérivée puisse être utilisé pour la recherche. Cependant, un aspect important du processus concerne l'importance relative des termes entre eux dans la requête. Ce concept est relatif à la pondération des termes au sein de la requête. Ce type d'information peut être donné par l'utilisateur, obtenu en utilisant la mesure *idf* ou simplement calculé en utilisant l'information de fréquence. Nous avons choisi de modéliser cette connaissance sous un corpus d'évidence associé à une fonction de masse $m^T[CO]$. Cette allocation est extraite de la requête par une distribution de possibilité sur l'ensemble T des termes dans la collection. Pour obtenir la fonction de masse qui tienne compte de cette information et du modèle précédemment décrit (relatif à la modélisation de l'incertitude quant au contenu d'un élément e), il nous faut combiner ces deux corpus d'évidence³. Ceci conduit à :

$$\mathbf{m}^T[e] = m^T[e] \oplus m^T[CO]. \quad (18)$$

Avec cette formulation, nous obtenons une fonction de masse $\mathbf{m}^T[e]$ dont les éléments focaux sont des sous-ensembles $A \subseteq T$ qui peuvent être des disjonctions d'éléments. De plus, si l'on ne veut pas prendre en compte cette information, il suffit d'initialiser la fonction de telle sorte qu'elle soit l'élément neutre de la règle de Dempster soit

2. En anglais, le terme 'topic statement' est utilisé.

3. Les lettres en gras sont utilisées pour les fonctions de masse obtenues après combinaison avec la requête.

$m^T[CO](T) = 1$. Pour les requêtes basées sur le contenu, nous utilisons la transformation pignistique, outil classique de la théorie des fonctions de croyance comme fonctions de recherche.

Ainsi, une requête notée $Q \subseteq T$ correspond à un ensemble de termes. Une fonction pignistique $BetP_{\mathbf{m}^T[e]}$ est calculée à partir de la fonction de masse $\mathbf{m}^T[e]$ qui modélise la représentation du document par :

$$BetP_{\mathbf{m}^T[e]}(Q) = \sum_{t \in Q} \sum_{A \subseteq T | t \in A} \frac{\mathbf{m}^T[e](A)}{|A|} \quad \forall Q \subseteq T. \quad (19)$$

La valeur de la fonction $BetP_{\mathbf{m}^T[e]}(\cdot)$ quantifie la pertinence d'un élément e par rapport à la requête Q . La première somme provient du fait qu'une requête Q contient un certain nombre de termes t tandis que la seconde correspond à la somme des masses attribuées aux éléments focaux dont l'intersection avec le terme t est non vide. Cette fonction est utilisée pour le classement des documents inclus dans la collection vis-à-vis de la requête. Bien que d'autres fonctions puissent être utilisées pour cette tâche⁴, nous avons choisi d'utiliser la transformation pignistique pour ses principales propriétés (linéarité, monotonie, ...).

5.2. Requête basée sur le Contenu et la Structure

Les requêtes basées sur le contenu et la structure font référence à la structure du document XML. Puisque ces requêtes restreignent leur champ d'investigation à des contextes particuliers (certains types de structure par exemple), il nous faut utiliser la connaissance modélisée par la fonction de masse $m^T[e, m^Y]$ qui dépend du contenu mais aussi de la structure. Pour prendre en compte l'importance des termes dans la requête, une fonction de masse, notée $m^T[CAS]$ définie sur l'ensemble des termes T peut être utilisée en relation avec la requête comme nous l'avons défini précédemment. En utilisant la règle de combinaison de Dempster, on obtient une représentation agrégée de l'incertitude relative au document et à la requête sous la forme :

$$\mathbf{m}^T[e, \mathbf{m}^Y] = m^T[e, m^Y] \oplus m^T[CAS]. \quad (20)$$

Comme pour les requêtes basées sur le contenu, nous utilisons la transformation pignistique comme fonction de recherche. Ainsi, la fonction $BetP_{\mathbf{m}^T[e, \mathbf{m}^Y]}$ est calculée à partir de $\mathbf{m}^T[e, \mathbf{m}^Y]$ pour chaque élément e par :

$$BetP_{\mathbf{m}^T[e, \mathbf{m}^Y]}(Q) = \sum_{t \in Q} \sum_{A \subseteq T | t \in A} \frac{\mathbf{m}^T[e, \mathbf{m}^Y](A)}{|A|} \quad \forall Q \subseteq T. \quad (21)$$

4. Dans [LAL 97], la pertinence d'un objet par rapport à une requête est évaluée par la fonction de croyance *bel* dans la phase de recherche.

6. Exemple illustratif

La méthodologie proposée est illustrée dans cette section au travers d'un exemple mettant en scène le document présenté initialement. La DTD associée à ce document nous permet de définir l'ensemble des éléments type $Y = \{CD, SONG, VERSE, CHORUS\}$. L'ensemble des éléments peut être déduit du document tel que $E = \{CD, SONG1, VERSE1, CHORUS1, SONG2, VERSE2a, VERSE2b, CHORUS2\}$. Soit un ensemble de termes T qui contient uniquement quatre index qui sont respectivement $T = \{'love', 'heart', 'peace', 'war'\}$. Les relations parent-fils entre les éléments peuvent également être déduites de la DTD associée au document XML. A partir de ce document et pour chaque élément e dans E , une fonction de masse $m^T[e]$ définie sur T est tout d'abord calculée. Le tableau 2 présente ces différentes fonctions de masse. En observant ce tableau, on peut remarquer que le terme $t_1 = 'love'$ indexe l'élément

$m^T[e](.)$	$t_1 = 'love'$	$t_2 = 'heart'$	$t_3 = 'peace'$	$t_4 = 'war'$	T
CD	0.300	0.100	0.300	0.100	0.200
SONG1	0.000	0.000	0.600	0.300	0.100
CHORUS1	0.000	0.000	0.800	0.000	0.200
VERSE1	0.000	0.000	0.600	0.200	0.200
SONG2	0.500	0.200	0.200	0.000	0.100
CHORUS2	0.600	0.000	0.000	0.000	0.400
VERSE2a	0.600	0.300	0.000	0.000	0.100
VERSE2b	0.600	0.000	0.300	0.000	0.100

Tableau 2. Fonctions de masse initiales déduites d'une collection de documents.

SONG2 tandis que le terme $t_3 = 'peace'$ a plutôt tendance à indexer l'élément SONG1.

A partir des fonctions de masse initiales, il faut calculer les masses affaiblies par rapport aux coefficients α_e qui sont définis pour tous les éléments de type SONG par $\alpha = 1$, pour tous les éléments de type VERSE, $\alpha = 0.9$ et $\alpha = 0.5$ pour tous les éléments de type CHORUS. Finalement, $m^T[children(e)](.)$ est obtenue en combinant les masses $\tilde{m}^T[e](.)$ qui formalisent la contribution des éléments-fils. En combinant avec la fonction de masse de l'élément lui-même, on obtient pour chaque élément e les fonctions de masse $m^T[e, children(e)]$ qui sont illustrées au tableau 3. La dernière colonne de ce tableau correspond aux valeurs allouées à la masse sur l'ensemble T qui quantifient l'incertitude quand à la connaissance sur l'indexation d'un élément du document. La pertinence de chaque élément contenu dans le document est évaluée en utilisant la transformation pignistique calculée avec l'équation (19). Les résultats sont présentés au tableau 4 pour six requêtes différentes. Sur ce tableau, les chiffres en gras mettent en évidence les éléments retrouvés par le système en utilisant la contribution de chaque élément-fils. Les astérisques sont quant à elles utilisées lorsqu'on ne prend pas en compte la contribution des éléments-fils (à partir des fonctions de masse du tableau 2). Comme on pouvait le prévoir, si l'on ne prend pas en compte la contribution des éléments-fils, les éléments retrouvés sont plus granulaires mis à part pour la

$m^T[e, children(e)]$	$t_1 = \text{'love'}$	$t_2 = \text{'heart'}$	$t_3 = \text{'peace'}$	$t_4 = \text{'war'}$	T
CD	0.325	0.016	0.641	0.014	0.002
SONG1	0.000	0.000	0.834	0.140	0.025
CHORUS1	0.000	0.000	0.400	0.000	0.600
VERSE1	0.000	0.000	0.540	0.180	0.280
SONG2	0.902	0.045	0.045	0.000	0.007
CHORUS2	0.300	0.000	0.000	0.000	0.700
VERSE2a	0.540	0.270	0.000	0.000	0.190
VERSE2b	0.540	0.000	0.270	0.000	0.190

Tableau 3. Fonctions de masse obtenues après la combinaison des masses qui formalisent la contribution des éléments-fils.

$BetP_{m^T[e]}$	Q_1	Q_2	Q_3	Q_4	Q'	Q''
CD	0.326	0.017	0.642	0.015	0.807	0.762
SONG1	0.006	0.006	0.841	0.147*	0.194	0.991*
CHORUS1	0.150	0.150	0.550*	0.150	0.890	0.862
VERSE1	0.070	0.070	0.610	0.250	0.755	0.922
SONG2	0.904	0.047	0.047	0.002	0.993	0.090
CHORUS2	0.475*	0.175	0.175	0.175	0.962	0.723
VERSE2a	0.587	0.317*	0.048	0.048	0.988*	0.358
VERSE2b	0.587	0.048	0.317	0.048	0.943	0.595

Tableau 4. Pertinence $BetP_{m^T[e]}(\cdot)$ des éléments par rapport aux Requêtes basées sur le Contenu. La requête Q_i correspond au terme t_i , $Q' = \{t_1, t_2\}$, $Q'' = \{t_3, t_4\}$.

requête Q_4 où le terme $t_4 = \text{'war'}$ n'apparaît pas réellement dans un élément-fils. En donnant plus de poids au terme $t_4 = \text{'war'}$, ceci a pour effet d'augmenter la valeur de pertinence pour l'élément VERSE1 permettant ainsi de conclure que la pondération des termes s'avère efficace dans ce contexte.

Nous illustrons maintenant le cas d'une Requête basée sur le Contenu et la Structure. Le tableau 5 présente les fonctions de masse m^Y correspondant à la structure du document. La fonction m_1^Y est obtenue à partir de la fonction *typew* en transformant la distribution de possibilité en une fonction de croyance. Les masses m_2^Y sont supposées données par l'utilisateur de telle manière à ce que le système accorde plus d'importance aux éléments de type CHORUS et VERSE. La dernière colonne de ce tableau illustre la fonction obtenue en combinant m_1^Y et m_2^Y par la règle de Dempster. Comme précédemment, le tableau 6 illustre la fonction de recherche calculée à partir de l'équation (21) pour les différentes requêtes. Ici, nous nous plaçons dans le même contexte que pour la recherche basée contenu (même requête, même pondération des termes) en imposant une contrainte sur les éléments cibles retournés par le système. Puisque la structure des éléments cible recherchés sont du type CHORUS et/ou VERSE,

.	$typew(\cdot)$	$m_1^Y(\cdot)$	$m_2^Y(\cdot)$	$m^Y(\cdot)$
CD	0.30	0.00	0.00	0.00
SONG	0.50	0.00	0.00	0.00
CHORUS	1.00	0.00	0.00	0.30
VERSE	0.70	0.30	0.00	0.00
CHORUS,VERSE	nd	0.20	0.90	0.65
SONG,CHORUS,VERSE	nd	0.20	0.00	0.02
Y	nd	0.30	0.10	0.03

Tableau 5. Fonctions de masse m^Y représentant la structure du document définie sur Y . La fonction m_1^Y est déduite de $typew$; m_2^Y est donnée par l'utilisateur. La dernière colonne correspond à la combinaison $m_1^Y \oplus m_2^Y$. (nd = valeur non définie)

$BetP_{m^T[e,m^Y]}$	Q_1	Q_2	Q_3	Q_4	Q'	Q''
CD	0.002	0.000	0.005	0.000	0.002	0.003
SONG1	0.000	0.000	0.009	0.002	0.001	0.006
CHORUS1	0.057	0.144	0.324	0.253	0.268	0.297
VERSE1	0.015	0.038	0.172	0.140	0.120	0.160
SONG2	0.051	0.006	0.003	0.000	0.006	0.002
CHORUS2	0.229	0.212	0.224	0.372	0.290	0.280
VERSE2a	0.405	0.549	0.087	0.144	0.158	0.108
VERSE2b	0.238	0.048	0.174	0.085	0.151	0.140

Tableau 6. Pertinence $BetP_{m^T[e,m^Y]}(\cdot)$ pour chaque élément par rapport à différentes requêtes. La requête Q_i correspond au terme t_i , $Q' = \{t_1, t_2\}$, $Q'' = \{t_3, t_4\}$.

il apparaît clairement (nombres en gras) dans le tableau 6 qu'il permette d'obtenir un bon classement à partir de la fonction de pertinence.

7. Conclusion

Dans cet article, un modèle théorique pour la recherche de documents XML basé sur la théorie des fonctions de croyance a été présenté. Ce cadre formel est utilisé pour représenter l'incertitude sur le contenu et la structure des documents. Il permet en outre la recherche de documents à partir de requêtes autorisant des conditions structurales. Un exemple d'une telle stratégie de recherche a été présenté et permet d'illustrer l'efficacité de la méthode.

Les travaux futurs concernent l'évaluation de notre approche sur le jeu de données INEX [BLA 03b]. Cependant, une des principales difficultés est l'ajustement des valeurs des différents paramètres liés aux incertitudes en particulier ceux liés à la structure. Même si l'approche semble flexible, les performances peuvent être moyennes si

un important effort n'est pas mené quand à l'ajustement correct de ces valeurs. Nous sommes en train de travailler sur ces aspects. Un autre point concerne les contraintes sur le contenu intrinsèque des requêtes.

8. Bibliographie

- [BAE 99] BAEZA-YATES R., RIBEIRO-NETO B., *Modern Information Retrieval*, Addison Wesley, 1999.
- [BER 02] BERGLUND A., BOAG S., CHAMBERLIN D., FERNANDEZ M., KAY M., ROBIE J., SIMEON J., « XML Path Language (XPath) 2.0 », W3C Working Draft, Novembre 2002, <http://www.w3.org/TR/xpath20>.
- [BLA 03a] BLANKEN H., T. GRABS R. S., WEIKUM G., Eds., *Intelligent Search on XML*, Springer-Verlag, 2003.
- [BLA 03b] BLANKEN H., GRABS T., SCHEK H., SCHENKEL R., WEIKUM G., Eds., *The INEX Evaluation Initiative*, Springer, 2003.
- [BOA 02] BOAG S., CHAMBERLIN D., FERNANDEZ M., FLORESCU D., ROBIE J., SIMEON J., « XQuery : An XML Query Language », W3C Working Draft, 2002, <http://www.w3.org/TR/XQuery>.
- [FUH 01] FUHR N., GROSSJOHANN K., « XIRQL : A query language for information retrieval in XML documents », *ACM SIGIR Conference on Research and Development in Information Retrieval*, New Orleans, USA, Août 2001.
- [FUH 03] FUHR N., GOEVERT N., KAZAI G., LALMAS M., Eds., *INEX : Evaluation Initiative for XML retrieval - INEX 2002 Workshop Proceedings*, DELOS Workshop, 2003.
- [KAM 03] KAMPS J., MARX M., DE RIJKE M., SIGURBJORNSSON B., « XML retrieval : What to retrieve ? », *ACM SIGIR Conference on Research and Development in Information Retrieval*, 2003, p. 409-410.
- [LAL 97] LALMAS M., « Dempster-Shafer's theory of evidence applied to structured documents : modelling uncertainty », *ACM SIGIR Conference on Research and Development in Information Retrieval*, 1997, p. 110-118.
- [RÖL 02] RÖLLEKE T., LALMAS M., KAZAI G., RUTHVEN I., QUICKER S., « The Accessibility Dimension for Structured Document Retrieval », *BCS-IRSG European Conference on Information Retrieval (ECIR)*, Glasgow, Mars 2002.
- [SHA 76] SHAFER G., *A Mathematical Theory of Evidence*, Princeton University Press, 1976.
- [SME 94] SMETS P., KENNES R., « The Transferable Belief Model », *Artificial Intelligence*, vol. 66, n° 2, 1994, p. 191-234.
- [THE 02] THEOBALD A., WEIKUM G., « The index-based XXL search engine for querying XML data with relevance ranking », *Advances in Database Thechnology - EDBT 2002, 8th International on Extending Database Technology*, Prague, Czech Republic, Mars 2002, p. 477-495.
- [WIL 94] WILKINSON R., « Effective Retrieval of Structured Documents », *ACM-SIGIR Conference on Research and Development in Information Retrieval*, Dublin, Irlande, 1994, p. 311-317.