Recherche d'information orientée contenu dans les documents XML par agrégation partielle des sources de pertinence

BAL kamal, NOUALI Omar

Centre de recherche sur l'information scientifique et technique (CERIST)
3, rue fréres Aissou, Ben Aknoun, Alger
k_bal@esi.dz, onouali@cerist.dz

1. Introduction

La recherche d'information (RI) orientée contenu dans les documents semi-structurés de type XML met en relation un besoin en information exprimé sous forme d'une requête sur le contenu recherché (liste de mots-clés) et une collection de document XML. Le système de recherche doit répondre en retournant non pas des documents entiers, mais juste des fragments de documents (des éléments XML) pertinents. Les éléments XML à restituer ne doivent pas seulement contenir l'information pertinente mais doivent être aussi d'un bon niveau de granularité. C'est-à-dire des éléments spécifiques et exhaustifs. La coexistence de l'information de structure et de contenu dans les documents XML et les spécificités liées à la recherche d'information dans ces documents font qu'une multitude de sources de pertinence hétérogènes et ayant des échelles de valeurs très variables peuvent être considérées dans la sélection des éléments pertinents et dans leur classement. Nous proposons une approche de recherche d'information orientée contenu dans les documents XML où le processus de recherche est guidé plus par la comparaison des éléments XML entre eux que par l'estimation de leurs scores de pertinence.

2. Agrégation partielle des sources de pertinence pour la RI XML

Nous modélisons le problème de la RI orientée contenu dans les documents XML comme un problème d'analyse multicritères. Les éléments XML constituent l'ensemble des alternatives possibles, les sources de pertinences constituent les différents critères d'évaluation et le problème de décision est un problème de rangement où les éléments XML doivent être rangés de plus au moins pertinent.

Soit : $E = \{e_1, e_2, \ldots, e_n\}$ l'ensemble des éléments XML, et

$C = \{C_1, C_2, \ldots, C_m\}$ l'ensemble des sources (critères) de pertinence

Chaque critère C_j sera modélisé par une fonction de même nom $C_j()$ donnant pour chaque élément XML *e,* sa performance par rapport à ce critère $C_j(e)$.

Une relation R_j sera associée à chaque critère de pertinence. Cette relation permettra de comparer chaque paire d'éléments XML selon ce critère. Ces relations doivent prendre en compte l'imprécision et l'arbitraire qui peuvent caractériser l'évaluation des différentes sources de pertinence. Une différence minime entre les performances de deux éléments XML sur un critère de pertinence ne doit pas obligatoirement les discriminer. Un seuil de d'indifférence q_j et un seuil de préférence p_j seront associés à chaque critère pour modéliser des relations de préférence graduelle (floues) comme suit :

$R_j(e_1,e_2) = 1$ si $C_j(e_1) - C_j(e_2) \geq p_j$

$R_j(e_1,e_2) = 0$ si $C_j(e_1) - C_j(e_2) \leq q_j$

$R_j(e_1,e_2) \in]0, 1[$ si $q_j < C_j(e_1) - C_j(e_2) < p_j$

Ces relation de préférence monocritères seront ensuite agrégées en une seule relation globale permettant de comparer chaque paire d'éléments XML sur l'ensemble des critères. ELECTRE III est une méthode de surclassement permettant de construire cette relation de préférence globale appelée relation de surclassement. ELECTREIII construit une relation de surclassement floue (notée S) caractérisée par un degré de crédibilité associé à chaque hypothèse de surclassement comme suit :

$S(e_1,e_2) = 1$ si tous les critères sont en faveur de « e_1 surclasse e_2 »

$S(e_1,e_2) = 0$ si aucun critère n'est en faveur de « e_1 surclasse e_2 »

$S(e_1,e_2) \in]0, 1[$ si quelque critères sont en faveur de « e_1 surclasse e_2 »

ELECTRE III exploite la relation de surclassement pour classer les éléments XML. Deux prés ordres totaux sont produits via deux procédures de distillation : La distillation descendante produit le premier pré ordre en retirant à chaque itération, de l'ensemble des éléments E, l'élément XML ayant la plus grande qualification (Q(e)), et en le classant en haut de la liste des résultats. La distillation ascendante retire à chaque itération l'élément ayant la plus faible qualification et le classe en bat de la liste des résultats. L'intersection des deux prés ordres donnera l'ordre final, avec :

$$Q(e) = \sum_{e'} S(e,e') - \sum_{e'} S(e',e) \qquad [1]$$

Un élément XML sera considéré plus pertinent qu'un autre élément si une majorité de critères de pertinence lui sont favorables qans qu'il y est de critères réfutant largement ce jugement.

L'intérêt de l'approche proposée réside dans la possibilité de considérer une multitude de sources de pertinence hétérogènes dans le processus de recherche et dans la prise en compte de l'imprécision dans l'évaluation des sources de pertinence.