Titre : Appariement sémantique de documents XML

Mots-clés : recherche d'information, XML, théorie des graphes, documents hétérogènes, appariement sémantique

Ce sujet de thèse rentre dans le contexte de la recherche d'information, et s'intéresse particulièrement à la recherche d'information dans les documents semi-structurés de type XML. La problématique engendrée par ce type de document est liée à la nature de leur contenu. En effet, comme ces documents comportent de l'information (du texte) et des contraintes structurelles (des balises), ils ne peuvent pas être efficacement exploités par les techniques classiques de RI, qui considèrent le document comme un granule d'information indivisible. Le défi à relever est alors d'arriver à identifier automatiquement l'unité d'information, en l'occurrence un élément du document XML, répondant à la requête de l'utilisateur.

De nombreuses approches ont été proposées dans la littérature pour permettre de renvoyer aux utilisateurs ces unités pertinentes [FLMK 2005], [FLT 06], [FKLT 07]. Ces approches se sont appliquées à renvoyer des parties de documents répondant de manière spécifique et exhaustive au besoin en information de l'utilisateur, exprimé sous forme de requête. Des représentations à base de graphes ou plus particulièrement d'arbres sont souvent utilisées, mais la théorie des graphes sous-jacente est peu exploitée. Cette dernière ne permet en effet que des processus d'appariement documents-requêtes basés sur des propriétés de structure. Dans le cas de recherche dans des corpus de documents XML, il est nécessaire d'y adjoindre une représentation complémentaire dans un autre formalisme, ce qui permettrait de prendre en compte la sémantique des documents et rendrait l'appariement complexe et imprécis.

La théorie des graphes pourrait pourtant être une aide précieuse pour l'appariement de structures, aide d'autant plus nécessaire lorsque les documents à traiter font partie de corpus hétérogènes. L'hétérogénéité des documents peut en effet porter sur plusieurs points : leur *taille* ou leur *contenu*, mais aussi leur *structure*.

Dans le cas de collections formées de documents possédant des tailles et des contenus différents, les méthodes proposées dans la littérature ne s'appliquent pas de manière optimale. En effet, les évaluations de pertinence des éléments ne peuvent pas s'effectuer de la même manière quand un document fait quelques Ko et qu'il possède une unité sémantique (il traite d'un même thème, aussi généraliste soit-il) que lorsqu'il fait 300 Mo et qu'il est conçu comme un catalogue de données. Des méthodes de correspondance d'arbres doivent être développées, et une réflexion doit être menée sur le traitement parallèle des documents orientés données et des documents orientés contenu.

Considérons maintenant l'hétérogénéité structurelle. Une collection possède des structures hétérogènes lorsque les documents qui la composent suivent des DTDs différentes. Alors que les approches proposées dans la littérature pour l'interrogation de corpus possédant des documents suivant la même DTD cherchent à vérifier des correspondances syntaxiques entre les arbres de la requête et des documents, les approches pour les corpus hétérogènes doivent chercher quant à elles à vérifier des correspondances sémantiques et permettre l'interrogation générique des corpus : les conditions de structures exprimées par les utilisateurs dans la requête ne correspondent pas forcément exactement aux schémas ou DTD des documents présents dans le corpus, mais ces derniers pourraient pourtant être pertinents pour l'utilisateur.

Plusieurs pistes de recherches sont possibles. Une première solution est d'utiliser un lexique, un thésaurus ou une ontologie pour faire correspondre les conditions de structures exprimées dans la requête avec les types d'éléments effectivement présents dans la collection [ThWe 02].

D'autres approches, comme celle proposée par Denoyer et al. dans [DeWG 04] ou Abiteboul et al. dans [AMNP 04] visent à proposer un format médian dans lequel tous les documents du corpus (et

éventuellement les requêtes) peuvent être transformés pour ensuite appliquer des techniques traditionnelles de traitement des requêtes structurées. Tous les documents devront cependant être transformés selon cette structure générique, au risque de perdre quelque peu de la sémantique portée par leur structure.
Pour construire et interroger des collections XML hétérogènes, l'apprentissage automatique des relations entre les différents formats et des transformations entre les différents documents est un ainsi problème important. La classification, le clustering et la correspondance de structures sont donc des défis majeurs à relever pour la gestion des documents semi-structurés.

Les propositions effectuées dans le cadre de cette thèse visant donc à s'appuyer sur des outils de la théorie des graphes pour la RI structurée dans les corpus hétérogènes, pourront être évaluées grâce à la campagne d'évaluation INEX, qui fournit un cadre générique pour l'évaluation de la recherche d'information structurée : collections, requêtes, jugements de pertinence et métriques. Plus précisément, les tâches *hétérogènes* et de *document mining* permettront se positionner les algorithmes proposés par rapport à d'autres systèmes de l'état de l'art.

Ce sujet de thèse est financé dans le cadre du projet ANR AOC (Appariement d'Objets Complexes). Plus d'informations sur : http://www.irit.fr/AOC

[AMNP 04] : Abiteboul S., Manolescu I., Nguyen B., Prada N., *A test platform for the INEX heterogeneous track*, Proceedings of INEX 2004, Dagstuhl, Allemagne, 2004.

[DeWG 04] Denoyer L., Wisniewski G., Gallinari P., *Document Structure matching for heterogeneous corpora,* Proceedings of XML and IR workshop, SIGIR 2004, Sheffield, England, 2004.

[FLMK 05]: N. Fuhr, M. Lalmas, S. Malik and G. Kazai, *INEX 2005 Workshop Proceedings*, Dagstuhl, Germany, 2005.

[FLT 06]: Norbert Fuhr Mounia Lalmas Andrew Trotman (Eds.). *Comparative Evaluation of XML Information Retrieval Systems*. 5th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2006 Dagstuhl Castle, Germany, December 17-20, 2006. Revised and Selected Papers

[FKLT 07]: Norbert Fuhr, Jaap Kamps, Mounia Lalmas, Andrew Trotman (Eds.). *Focused Access to XML Documents*. 6th InternationalWorkshop of the Initiative for the Evaluation of XML Retrieval, INEX 2007. Dagstuhl Castle, Germany, December 17-19, 2007. Revised and Selected Papers

[SaBC 06] Sauvagnat K., Boughanem M., Chrisment C., *Answering content-and-structure-based queries on XML documents using relevance propagation*, Information Systems, Special Issue SPIRE 2004, volume 31, p. 621-635, janvier 2006.

[ThWe 02] Theobald A., Weikum G., *The Index-Based XXL Search Engine for Querying XML Data with Relevance Ranking*, EDBT 2002, 8th International Conference on Extending Database Technology, Prague, Czech Republic, p. 477-495, 2002.

Equipe de recherche : IRIT – SIG/RFI
Directeur de recherche : Mohand Boughanem
bougha@irit.fr
Tel : 05 61 55 74 16
Encadrement : Karen Pinel-Sauvagnat
sauvagna@irit.fr
Tel : 05 61 55 74 41