Sujet de thèse : Graphes et Appariement sémantique de documents XML

Doctorant : Mr. Lei NING

Directeur de thèse : Pr. Hamamache KHEDDOUCI

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

Cette thèse porte sur la recherche d'information dans des corpus de documents XML hétérogènes. Le développement d'outils automatisés permettant un accès efficace à cette quantité gigantesque d'information numérique issue de l'environnement Internet apparaît comme une nécessité. De nombreuses méthodes ont été proposées dans la littérature pour permettre de renvoyer aux utilisateurs de l'information pertinente issue de documents XML [Fuhr 2005]. Ces méthodes se sont appliquées à renvoyer des parties de documents répondant de manière spécifique et exhaustive au besoin en information de l'utilisateur. Elles n'ont cependant pas (ou peu) pris en compte l'hétérogénéité des corpus considérés. 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 cherchent quant à elles à vérifier des correspondances sémantiques. Des méthodes automatiques à base de graphes doivent être trouvées pour 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 [The 2002]. D'autres approches, comme celle proposée par Denoyer et al. dans [Den 2004] ou Abiteboul et al. dans [Abi 2004] 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, défis réalisables en coordonnant des techniques d'intelligence artificielle, des techniques de recherche d'information et des techniques d'appariement de graphes. Ces travaux de thèse s'appuieront sur le modèle XFIRM développé au sein de l'équipe SIG de l'IRIT, modèle

permettant d'effectuer des recherches flexibles dans des corpus de document semi-structurés. Le modèle de recherche est basé sur une méthode de propagation de la pertinence, ayant pour but de trouver les unités d'information les plus exhaustives et spécifiques répondant à une requête utilisateur, que celle-ci contienne ou non des conditions de structure. Les documents semi-structurés pouvant être représentés sous forme arborescente, et le but est alors de trouver les sous-arbres de taille minimale répondant à la requête.

Les recherches sur le contenu seul des documents sont effectuées en prenant en compte les importances diverses des feuilles des sous-arbres, et en plaçant ces derniers dans leur contexte, c'est à dire, en tenant compte de la pertinence du document. Les recherches portant à la fois sur le contenu et la structure des documents sont effectuées grâce à plusieurs propagations de pertinence dans l'arbre du document, et ce afin d'effectuer une correspondance vague entre l'arbre du document et l'arbre de la requête [Sau 2006]. Les propositions effectuées dans le cadre de cette thèse pourront être évaluées grâce à la campagne d'évaluation INEX [Fuhr 2005], 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.

Références

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

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

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

[Sau 2006] 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.

[The 2002] 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.

[Fuhr 2006]: N. Fuhr, M. Lalmas, A. Trotman, INEX 2006 Workshop Proceedings, Dagstuhl, Germany, 2006.