XFIRM: Recherche d'information dans des documents XML

1. Présentation

Ce prototype permet d'effectuer de la recherche d'information dans des collections de documents textes encodés au format XML. Il prend en entrée une requête utilisateur exprimant son besoin sur la collection de documents et retourne en réponse à cette requête une liste d'éléments XML triés par ordre décroissant de pertinence.

1.1. Types de graphes traités par le prototype

Arbres orientés et étiquetés.

Plus précisément : collection de documents textes encodés au format XML, requêtes avec conditions de contenu et structure également encodées au format XML.

1.2. Spécifications techniques

Notre prototype se base sur le moteur de recherche XFIRM décrit dans [1].

Le code est en Java (compilé avec JDK 1.6) et d'un certain nombre de procédure PL-SQL principalement utilisées pour l'indexation.

Les index se basent sur les algorithmes présentés dans [5].

1.3. Format entrée-sortie

Notre prototype utilise des requêtes CAS au format NEXI [2] issues des éditions 2005, 2010 et 2011 de la campagne d'évaluation INEX dont une présentation générale est faite dans [3] et [4].

Il retourne une suite d'éléments classés par degrés de pertinence dont le format est compatible avec EvalJ, l'outil d'évaluation d'INEX.

2. Algorithmes implantés pour la recherche

Les algorithmes implantés sont basés sur la distance d'édition entre arbres [7], avec des coûts d'insertion, suppression et renommage calculés en fonction de la DTD des documents [8]. Afin de réduire le temps d'exécution, des résumés d'arbres peuvent être utilisés [6].

3. Evaluation et résultats

3.1. *Tâche d'évaluation et collection*

La campagne de référence pour l'évaluation de la recherche d'information dans des documents XML est INEX. Nous nous sommes intéressés plus particulièrement :

- aux requêtes de type « Content and Structure », exprimables sous forme d'arbre
- à deux collections de documents, fournies en 2005 et 2010 : la collection IEEE (environ 15000 articles scientifiques IEEE balisés au format XML) ainsi que la collection Datacentric (données du site Internet IMDB, 4.4 millions de documents).

3.2. Résultats

Les résultats sont décrits dans [6] [7] [8].

4. Bibliographie

[1] Karen Sauvagnat. Modèle flexible pour la recherche d'information dans des corpus de documents semi-structurés. Thèse de doctorat, Université Paul Sabatier, juin 2005.

[2] A. Trotman. Narrowed Extended XPath I (NEXI). In Proceedings of the INEX 2004 Workshop, pages 16-40. Springer-Verlag GmbH, 2004.

[3] G. Kazai and M. Lalmas. INEX 2005 evaluation measures. In Proceedings of the Fourth Workshop of the INitiative for the Evaluation of XML Retrieval (INEX 2005), volume 3977 of Lecture Notes in Computer Science, pages 16-29. Springer Verlag, 2006.

[4] Proceedings of the INitiative for the Evaluation of XML retrieval workshop (INEX 2010). Vught (Netherlands), 2010.

[5] Y.Lee, S.Yoo et K.Yoon « Index structures for structured documents » In proceedings of the ACM Workshop on XML and IR, Bethesda, pages 91-99, 1996.

[6] Cyril Laitang, Mohand Boughanem, Karen Pinel-Sauvagnat. XML Information Retrieval through Tree Edit Distance and Structural Summaries. Dans : Asia Information Retrieval Society Conference (AIRS 2011), Dubai, United Atab Emirates, 18/12/2011-20/12/2011, Springer, p. 73-83, décembre 2011.

[7] Cyril Laitang, Karen Pinel-Sauvagnat. Utilisation de la théorie des graphes et de la distance d'édition pour la recherche d'information sur documents XML. Dans : Conférence francophone en Recherche d'Information et Applications (CORIA 2011), Avignon, 16/03/2011-18/03/2011, p. 349-364, mars 2011.

[8] Cyril Laitang, Karen Pinel-Sauvagnat, Mohand Boughanem. Coûts de distance d'édition pour la Recherche d'Information XML. Dans : Conférence francophone en Recherche d'Information et Applications (CORIA 2012), Bordeaux, 21/03/2012-23/03/2012.