Proposition de recherche doctorale

Développement d'algorithmes et d'outils pour le support de l'archivage du Web

Mots clés :

- **Directeur de thèse** : Stéphane Gançarski
- **Co-encadrant(s)** :
- **Unité de recherche** : Laboratoire d'informatique de Paris 6
- **Ecole doctorale** : École Doctorale Informatique, Télécommunications, Électronique de Paris
- **Domaine scientifique principal**: Divers

Résumé du projet de recherche (Langue 1)

Le besoin de préserver l'information venant du Web et les moyens de gérer des archives de documents venant du Web sont des problèmes largement étudiés ces dernières années, et la plupart des pays ont initiés des projets de grande ampleur à des fins de préservation du patrimoine numérique mais aussi en vue de constituer des dépôts légaux. [9] fait un survey d'une grande partie des travaux de recherche du domaine. Dans la dernière décennie, avec l'émergence du Web comme source d'information à large échelle, beaucoup de gouvernements (et d'organismes internationaux) ont initiés des projets sur le sujet, principalement au travers des bibliothèques et instituts d'archivage nationaux. Ainsi, le portail du International Internet Preservation Consortium (IIPC) présente les différentes initiatives nationales existantes en vue de constituer une archive numérique [8]. Le projet Internet Archive a développé une infrastructure pour l'archivage de documents du Web et leur archive contient déjà 500 TO en versions de pages Web et 500 autres TO pour les documents associés (images, vidéos, etc.) [6]. Ce projet, initié en 1996, propose un accès public à son archive depuis 2000. Les différentes phases ou tâches concernées par l'archivage sont : * sélection des pages à archiver (définition du périmètre du corpus) * capture régulière du contenu des pages sélectionnées stockage et * indexation des versions de page capturées recherche d'information et * interrogation de l'archive préservation de l'archive constituée Pour chacune de ces phases, il est nécessaire de définir des stratégies efficaces, des algorithmes et des outils spécifiques capable de passer à l'échelle. Certains travaux se concentrent sur la capture des informations, effectuée régulièrement, voire périodiquement. Les pages sélectionnées sont visités, téléchargées puis stockées. Généralement, cette tâche est effectuée de manière automatique à l'aide de robots (crawlers) capable de visiter un site et capturer toutes les pages, ou simplement celles accessibles par un chemin borné depuis la racine du site. Cependant, afin de limiter l'utilisation des ressources, il est nécessaire de bien calculer à quel moment le robot doit revisiter une page pour éviter de capturer de l'information redondante ou peu nécessaire (peu de différence avec la version précédente). Ce problème est difficile puisque les modifications apportées sur les sites Web ne sont pas connues du côté de l'archive et il faut donc les prévoir en se basant sur les précédentes captures. Plusieurs travaux se sont penchés sur le sujet. [11] présente le système AOLAP (Austrian On-Line Processing Module) utilisant des techniques d'analyse des entrepôts de données, en incorporant dans les métadonnées des éléments provenant du service Whois. D'autres approches [12, 13, 14] se focalisent sur la modélisation et l'estimation de la fréquence des changements pour chaque page Web. Ils proposent des estimateurs de fréquence de changement (basés sur le modèle de Poisson pour la plupart) afin de prévoir le meilleur comportement possible du robot dédié à chaque page. Les travaux menés par le LIP6 dans le domaine de l'archivage du Web couvrent l'ensemble du processus par une approche originale, depuis l'analyse de l'aspect visuel des pages (car c'est celui que voient les utilisateurs, et celui sur lequel est basé la notion de dépôt légal) jusqu'à l'ordonnancement des crawlers, l'indexation des versions de page et l'exploitation de l'archive par des technique de recherche d'information temporelle incomplète [1,2,3,4]. L'un des points clés de cette approche est de définir des mesures, calculées en comparant l'aspect visuel (par segmentation) des versions de page successivement captées. La segmentation, effectuée à l'aide d'une version étendue du logiciel VIPS proposé par Cai et al. [15], consiste a décomposer celle-ci en bloc sémantiques, organisés en une hiérarchie (document VI-XML). Les documents VI-XML sont ensuite passés à VI-Diff, un algorithme original permettant de comparer deux versions successives de page segmentées et de détecter ainsi les différences entre les deux versions. On affecte à chaque bloc une importance relative liée à sa position dans la page et obtenue par apprentissage, ce qui permet de calculer une importance à chaque différence détectée, et ainsi de calculer, par agrégation, l'importance du changement entre deux versions. Cette différence, ainsi que le temps séparant les captures respectives des deux versions comparées, permet d'estimer le comportement dynamique de la page, et de prévoir le meilleur moment pour aller la revisiter, en fonction des limitations de ressources (les pages dont les changements importants sont les plus fréquents seront visitées plus souvent). Si le logiciel VIPS utilisé pour la segmentation a donné de bons résultats dans les premières expérimentations faites au LIP6, nous nous proposons, dans le cadre de la thèse, de fournir une alternative sous la forme d'une implémentation en logiciel libre qui puisse être utilisée sur les plateformes du LIP6 comme du CCPD-UCV. Ce logiciel alternatif, outre qu'il lèvera les limitations en terme de liberté d'utilisation, sera optimisé pour l'utilisation dans le système d'archivage, contrairement à VIPS qui est plus générique et donc moins performant. De plus, ce logiciel pourra être paralléliser afin d'améliorer le temps de traitement, ce qui est crucial puisque pour bien faire, un maximum de versions de page capturées devrait être analysé. Par exemple, plus de versions sont analysées, plus précis sera l'ordonnancement des crawlers. Enfin, une page Web pouvant être considérée comme une image, on peut utiliser des techniques de traitement d'image afin de rendre la segmentation encore plus précise. Une autre tâche d'archivage du Web concerne le stockage des pages et l'interrogation/recherche d'information dans l'archive constituée. Pour cela, on étudiera les travaux existant pour le stockage des pages dans des formats spécifiques [16, 17]. Concernant l'indexation des pages archivées, on se basera sur un format existant, comme le format DAFF mis au point par l'INA (partenaire du LIP6 dans le projet ANR Cartec), qu'on enrichira à l'aide de méta-information comme l'importance des changements (voir plus haut) ou toute autre information pertinente permettant de faciliter et accélérer la recherche d'information dans l'archive. Le cadre de ces travaux sera le projet européen Scape (FP7), projet sur la préservation numérique auquel participe l'équipe d'accueil de la thèse. Références bibliographiques :

[1] Myriam Ben Saad, Zeynep Pehlivan, Stéphane Gançarski. Coherence-oriented Crawling and Navigation for Web Archives using Patterns. In TPDL '11: Proceedings of the 15th International Conference on Theory and Practice of Digital Libraries (formerly ECDL: European Conference on Digital Libraries), Berlin, Germany, September, 2011 [2] Myriam Ben Saad, Stéphane Gançarski, Zeynep Pehlivan. A NovelWeb Archiving Approach based on Visual Pages Analysis. 9th InternationalWeb ArchivingWorkshop (Corfu 2009) http:// www.iwaw.net/09, collocated with ECDL09. [3] Saad, M. B. and Gançarski, S. 2010. Using visual pages analysis for optimizing web archiving. In Proceedings of the 2010 EDBT/ICDT Workshops (Lausanne, Switzerland, March 22 - 26, 2010). EDBT '10, vol. 426. ACM, New York, NY, 1-7. [4] Zeynep Pehlivan, Myriam Ben Saad, Stéphane Gançarski. Vi-DIFF: Understanding Web Pages Changes. 21 st International Conference DEXA 2010. 1-15, (Bilbao, Spain, August 30 – September 3, 2010) LNCS 6261. Spinger. ISBN 978-3-642-15363-1. [5] Andres Sanoja, Claudia León and Gustavo Torres. Lineamientos para la Construcción de un Archivo Histórico de la Información Digital producida en Venezuela. CLCAR 2010. Conferencia Latino Americana de Computación de Alto Rendimiento. August 25-28, 2010, Gramado, RS, Brazil. http://gppd.inf.ufrgs.br/clcar2010/program.html. [6] Masanès, Julien (Ed.). Web Archiving. Springer-Verlag. ISBN: 978-3-540-23338-1. 2006. [7] Ball, Alex. Web Archiving (version 1.1). Edinburgh, UK: Digital Curation Centre. 2010. [8] IIPC Web Sitee http://netpreserve.org/ [9]http://wwliver Witvoet and Andreas Aschenbrenner and Robert Bruckner. Putting the World Wide Web into a Data Warehouse: A DWH-Based Approach to Web Analysis. Database and Expert Systems Applications, International Workshop on. 2002, p: 822 [12] J. Cho and H. Garcia-Molina. The Evolution of the Web and Implications for an Incremental Crawler. In VLDB '00: Proceedings of the 26th International Conference on Very Large Data Bases, 2000. [13] J. Cho and H. Garcia-Molina. Estimating frequency of change. ACM Trans. Interet Technol., 3(3), 2003. [14] S. R. Singh. Estimating the rate of web page updates. In IJCAI, 2007. [15] D. Cai, S. Yu, J.-R. Wen, and W.-Y. Ma. VIPS: a Vision-based Page Segmentation Algorithm. Technical report, Microsoft Research, 2003. [16] W. Cathro. Development of a digital services architecture at the national library of Australia. EduCause, 2003. [17] D. Gomes, A. L. Santos, and M. J. Silva. Managing

duplicates in a web archive. In SAC’06: Proceedings of the 2006.