Le projet RESURGENCE :

RECOUVREMENT DE LA STRUCTURE DES DOCUMENTS ELECTRONIQUES

Emmanuel GIGUET, Nadine LUCAS, Catalina CHIRCU

GREYC UMR 6072
CNRS – Université de Caen Basse-Normandie – ENSICAEN

MOTS CLÉS :
Ingénierie Linguistique, Structure des documents

PRÉSENTATION :

Le projet que nous souhaitons présenter au cours de la session « Show&Tell » se situe à la frontière de l'ingénierie linguistique et de l'analyse d'image. Il concerne le recouvrement de la structure des documents électroniques. Ce projet fait l'objet d'un partenariat entre le laboratoire GREYC (CNRS UMR 6072) et la société MEMODATA, via un financement du Pôle régional bas-normand ITIC (Imagerie et Technologies de l'Information et de la Communication). Le projet touche à sa fin.

Le projet RESURGENCE traite le problème de l'accès au contenu textuel de documents de format hétérogène (e.g., PDF, Word, RTF, HTML, Postscript,…) ainsi que le problème connexe du recouvrement de leur structure logique (inférence des titres, sous-titres, sections, résumés, introductions, bibliographies, copyrights, …). Il porte plus particulièrement sur l'analyse des articles scientifiques et des articles journalistiques, quelle que soient leur langue et le format de document.

L'enjeu du projet RESURGENCE est de faciliter la constitution et l'analyse de corpus de documents hétérogènes, ainsi que de permettre la prise en compte de la structure dans les traitements informatiques, notamment pour la recherche d'informations. A terme, il s'agit également de faciliter la présentation d'un document sur différents supports (écran, papier, pda, …), d'autoriser des parcours ou des modes de lecture appropriés, par exemple, aux déficients visuels.

L'accès au contenu textuel des documents électroniques est encore aujourd'hui très limité. La structure du document est en effet très rarement exploitée par les logiciels actuels car elle est soit absente du format du document (i.e., pdf) soit partiellement codée dans un format propriétaire non ou peu documenté (i.e., doc). Pour permettre l'exploitation d'un contenu structuré, nous avons mis au point un système permettant d'inférer la structure logique. Il utilise un minimum de connaissance spécifique à une langue pour permettre son utilisation dans un environnement multilingue.

Nous présenterons et illustrerons la chaîne de traitement mise au point au laboratoire GREYC, qui à partir d'articles scientifiques publiés dans des formats variés et écrits dans les langues variées effectue une conversion dans un format pivot et réalise ensuite le recouvrement de la structure sous-jacente, Nous commenterons les techniques utilisées, les difficultés rencontrées et les perspectives.

CONTACT :

Emmanuel.Giguet@info.unicaen.fr, Nadine.Lucas@info.unicaen.fr