Application BASTRI
Fiches Equipes

MOSTRARE (SR0039GR)

modèles de structures arborescentes, apprentissage et extraction d'information
MOSTRARE (LINKS (SR0549UR) , MAGNET (SR0550PR))

Statut: Terminée

Responsable : Joachim Niehren (Par intérim)

Mots-clés de "A - Thèmes de recherche en Sciences du numérique - 2023" : *Aucun mot-clé.*

Mots-clés de "B - Autres sciences et domaines d'application - 2023" : *Aucun mot-clé.*

Domaine : Perception, cognition, interaction
Thème : Représentation et traitement des données et des connaissances

Période : 01/04/2004 -> 31/12/2012
Dates d'évaluation : 11/10/2011

Etablissement(s) de rattachement : U. LILLE 1 (USTL), CNRS, U. LILLE 3 (UCDG)
Laboratoire(s) partenaire(s) : LIFL (UMR8022)

CRI : Centre Inria de l'Université de Lille
Localisation : Centre Inria de l'Université de Lille
Code structure Inria : 101017-0

Numéro RNSR : 200418288R
N° de structure Inria: SR0039GR

Présentation

Le Web est désormais le plus grand entrepôt de données qui ait jamais existé. Cet entrepôt n'a pas de structure définie et les données sont hétérogènes et réparties. Cependant les utilisateurs veulent le considérer comme un système d'information que l'on puisse interroger facilement tout en obtenant des réponses pertinentes aux questions. Le besoin d'outils de recherche d'information et d'extraction d'information est donc essentiel. Les formats du Web évoluent avec l'apparition de XML et, peut-être l'apparition du Web sémantique.
L'objectif de l'équipe-projet est le développement de nouvelles techniques de recherche et d'extraction d'information utilisant la structure arborescente des documents. Les problèmes nouveaux que nous considérons sont :

- la **définition de modèles et d'algorithmes** pour des structures arborescentes adaptés à la tâche d'extraction d'information
- la **conception d'algorithmes d'apprentissage** artificiel utilisant les structures arborescentes des données et documents

Axes de recherche

- **Structures arborescentes et extraction d'information** Cet axe concerne l'étude des modèles et des algorithmes pour des données et documents possédant une structure arborescente avec pour objectif la tâche d'extraction d'information. Trois axes d'étude seront privilégiés :
  - le point de vue *automates* permettant de définir des classes de langages d'arbres et des classes d'algorithmes ;
  - le point de vue *logique* dont la correspondance avec les automates est bien connu.
  - le point de vue *contraintes* complémentaire des deux précédents.

  Pour chacun de ces points de vue, l'objectif est de définir des classes suffisamment expressives relativement au problème d'extraction d'information tout en conservant de bonne propriétés algorithmiques. En particulier, il s'agit de développer des "tree wrappers", programmes d'extraction d'information sur des données arborescentes.
- **Algorithmes d'apprentissage à l'aide de structures arborescentes** L'objectif est ici de développer de nouveaux algorithmes d'apprentissage utilisant la structure arborescente des données et documents. Des algorithmes de classification et de recherche d'information, ainsi que des algorithmes de construction de "tree wrappers" à partir d'exemples. Les techniques de combinaison de

Contact

- **Responsable :** Joachim Niehren
- **Tél :** + 33. 3. 5.9 .57. 7.8 .48
- **Secrétariat Tél :** + 33. 3. 5.9 .57. 7.8 .38

En savoir plus

- Site sur inria.fr
- Site du responsable
- Derniers Rapports d'Activité :

Documents sur la structure

- Intranet
- Privés

Décisions

- 4136 (02/06/2004) : création
- 5156 (10/10/2006) : prolongation
- 6249 (30/09/2008) : changement de rattachement
- 7026 (16/12/2009) : prolongation
- 8083 (27/07/2011) : cessation du responsable
- 8084 (27/07/2011) : nomination responsable
- 8929 (14/01/2013) : prolongation
- 9104 (14/01/2013) : fermeture

Localisation

- **Adresse postale :** Centre Inria de l'Université de Lille Parc Scientifique de la Haute Borne 40, avenue Halley Bât.A, Park Plaza 59650 Villeneuve d'Ascq France
- **Coordonnées GPS :** 50.606, 3.149

méthodes telles que le boosting, le co-training seront privilégiées.

Relations industrielles et internationales

- Industrielle : projet de collaboration avec XRCE - Xerox Grenoble, LIXTO - Vienna
- Institutionnelle : ACI Masse de données, AS DSTIC
- Scientifique internationale : Université, DFKI et MPI Saarebruck ; Vienne ; Utrecht ; NAIST (Japan) ; Trêves ; Barcelone ; Iasi.