Mostrare dans le projet WebContent (Lot 2, Extraction de contenu)

Fabien Torre

Équipe INRIA Futurs Mostrare
http://mostrare.lille.inria.fr/

11 juillet 2006

Équipe Mostrare

Fiche signalétique

- INRIA Futurs, créée en 2004 à Lille ;
- compétences en logique, automates et apprentissage ;
- membres de deux équipes du Laboratoire d'Informatique Fondamentale de Lille : STC (Lille 1) et GRAppA (Lille 3) ;
- une quinzaine de membres, un tiers de doctorants.

Objectifs de Mostrare

- Production automatique de requêtes et outils d'extraction d'information ;
- dans les documents semi-structurés ;
- par des techniques d'apprentissage automatique.

Axes de recherche

Modèles de structures arborescentes

- Point de vue *automates d'états finis* pour les langages d'arbres ;
- point de vue *logique* pour les langages d'arbres ;
- objectif : algorithmes adaptés à XML pour apprendre les requêtes d'extraction d'information.

Algorithmes d'*apprentissage* à partir de structures arborescentes

- Données : des documents XML ;
- tâches : clustering, classification, inférence de langages d'arbres, extraction et annotation.

Résultats

Modèles formels

- Automates pour arbres d'arité non bornée ;
- OcamlQuery : requêtes *n*-aires sur des documents XML ;
- résultats de complexité sur différents fragments XPath, requêtes dans les graphes.

Systèmes

- Squirrel : inférence d'automates d'arbres pour l'extraction d'information unaire dans des pages Web ;
- PaF : extraction *n*-aire par techniques de classification itérées.

Travaux en cours

Squirrel - Julien Carme - Aurélien Lemay

- Thèse de Julien Carme (Rémi Gilleron, Marc Tommasi et Joachim Niehren) ;
- Squirrel : inférence d'automates d'arbres pour l'extraction de relations unaires ;
- heuristique d'élagage pour les documents partiellement annotés ;
- interactif avec une interface utilisateur ;
- à venir : passage au n-aire par Aurélien Lemay.

Travaux en cours

Squirrel - Julien Carme - Aurélien Lemay

PaF - Patrick Marty

- Thèse de Patrick Marty (Rémi Gilleron, Marc Tommasi et Fabien Torre) ;
- identification d'organisations complexes des données ;
- problème d'extraction reformulé en apprentissage supervisé attributs-valeurs ;
- extraction de relations unaires et n-aires ;
- en cours et à venir : prise en charge des documents partiellement annotés, apprentissages interactif et actif.

Travaux en cours

Squirrel - Julien Carme - Aurélien Lemay

PaF - Patrick Marty

Tuareg et Suse - Laurent Candillier

- Thèse de Laurent Candillier (Isabelle Tellier et Fabien Torre)
- *subspace clustering*, combinaisons d'apprentissages supervisé et non supervisé ;
- proposition de codages attributs valeurs pour les documents XML ;
- réussite au challenge INEX.

Travaux en cours

Squirrel - Julien Carme - Aurélien Lemay

PaF - Patrick Marty

Tuareg et Suse - Laurent Candillier

XCRF - Florent Jousse

- Thèse de Florent Jousse (Rémi Gilleron, Marc Tommasi et Isabelle Tellier) ;
- *Conditionnal Random Fields* étendus aux arbres ;
- annotation automatique d'arbres (cas particulier : extraction d'informations).

Travaux en cours

Squirrel - Julien Carme - Aurélien Lemay

PaF - Patrick Marty

Tuareg et Suse - Laurent Candillier

XCRF - Florent Jousse

Mostrare et WebContent

Les personnes impliquées

- Laurent Candillier (systèmes Tuareg et Suse) ;
- Rémi Gilleron (responsable Mostrare)
- Florent Jousse (XCRF) ;
- Aurélien Lemay (Squirrel n-aire) ;
- Patrick Marty (système PaF) ;
- Isabelle Tellier ;
- Marc Tommasi ;
- Fabien Torre (contact WebContent).

Systèmes interactifs

Schéma interactif indépendant de l'algorithme d'apprentissage.

Scénario général

- L'utilisateur désigne dans une page web *quelques* éléments à extraire ;
- l'algorithme *infère* un automate qui propose des extractions dans la page initiale ou sur toute page de même type ;
- l'utilisateur corrige *certaines* des éventuelles erreurs.
- En *peu d'interactions*, on converge vers un wrapper qui réalise parfaitement la tâche.

Systèmes interactifs

Schéma interactif indépendant de l'algorithme d'apprentissage.

Scénario général

- L'utilisateur désigne dans une page web *quelques* éléments à extraire ;
- l'algorithme *infère* un automate qui propose des extractions dans la page initiale ou sur toute page de même type ;
- l'utilisateur corrige *certaines* des éventuelles erreurs.
- En *peu d'interactions*, on converge vers un wrapper qui réalise parfaitement la tâche.

Réalisations logicielles à venir

- Interface utilisateur pour l'interaction (extension Firefox) ;
- Web Service pour l'extraction.

Bilan Mostrare dans WebContent

Objectifs

- Exploiter les données issues d'outils de reporting ;
- reconnaître automatiquement des organisations complexes ;
- proposer des scénarios interactifs ;
- limiter les interactions avec l'utilisateur ;
- définir des stratégies d'apprentissage actif.

Apports

- Algorithmes d'apprentissage de wrappers ;
- algorithmes d'annotation automatique de documents XML ;
- interface utilisateur pour l'apprentissage interactif ;
- plate-forme d'extraction pouvant accueillir tout algorithme d'apprentissage de wrappers.