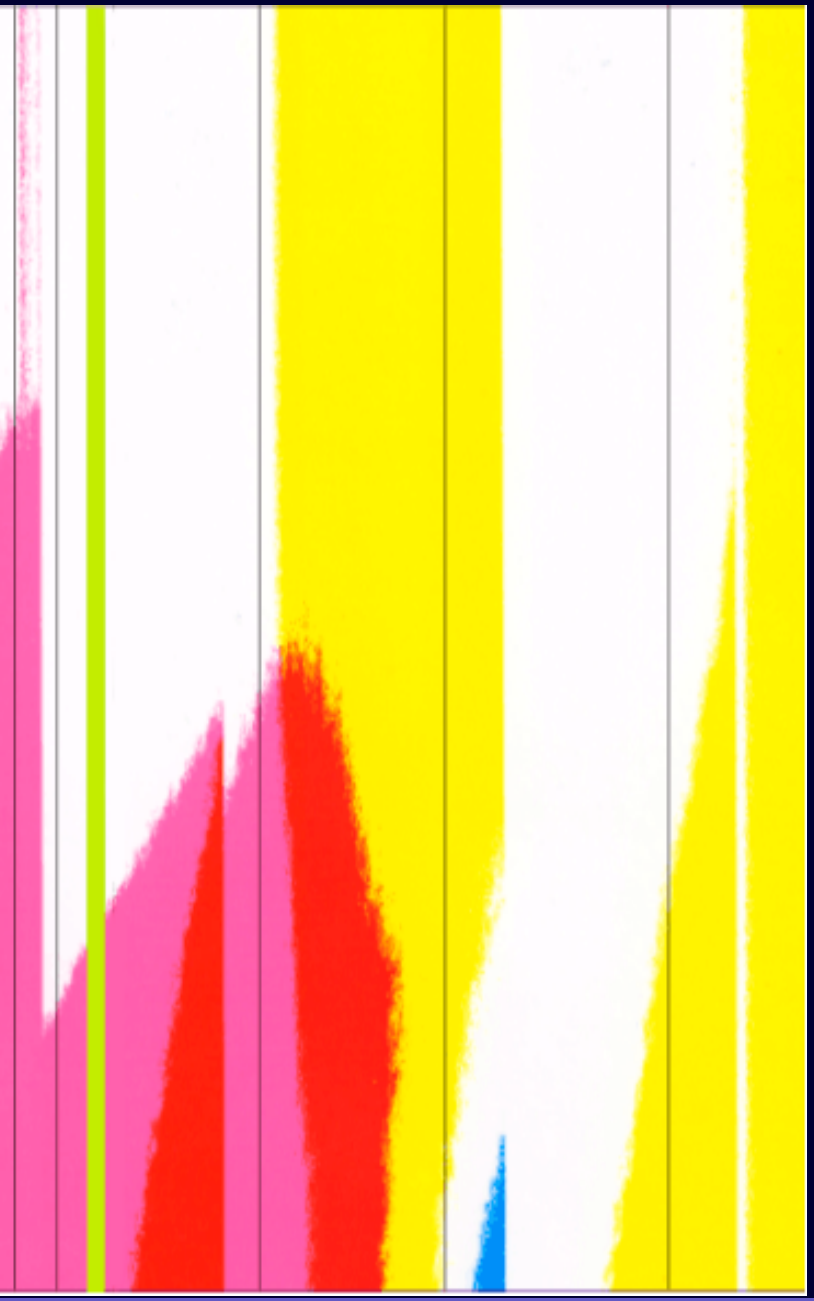

Mostrare dans le projet WebContent

Marc Tommasi

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

8 février 2011

Équipe Mostrare

Fiche signalétique

- Créée en 2004
- Regroupe des compétences en logique, automates et apprentissage.
- Aujourd'hui une dizaine de chercheurs et enseignants chercheurs

Objectifs de Mostrare

- génération 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 : classes d'algorithmes adaptés à `Xml` pour les requêtes d'extraction d'information.

Algorithmes d'apprentissage à l'aide de structures arborescentes

- Les données : des documents `Xml`
- Les tâches : classification, annotation, inférence de langages d'arbres

Réalisations

Modèles

- Automates pour arbres d'arité non bornée.
- OcamlQuery : requêtes n-aires sur des documents `Xml`.

Apprentissage

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

Focus : Inférence grammaticale et extraction

Situation

- l'utilisateur désigne dans une page web quelques éléments à extraire ;
- l'algorithme infère l'automate capable d'extraire tous les éléments à extraire, sur toutes les pages de même type.
- Le tout doit se faire en peu d'interactions.

Dans WebContent

Interactions

- Objectif : Limiter les interactions
- Outils : Développer des procédures d' apprentissage actif.

Reporting et tableaux de bord

- Objectif : Exploiter les données issues d'outils de reporting, d'aide à la décision. Découvrir automatiquement des structures en tableaux complexes (tables croisées, tournées, factorisées...)
- Outils : Extraction n-aire.

Dans WebContent

Annotations

- Objectifs : fournir des algorithmes d'annotation automatique de documents semi-structurés
- Outils : inférence statistique d'étiquetages.

Transformations

- Objectifs : favoriser l'intégration de données par des outils de transformation inférés à partir d'exemples.
- Outils : Inférence statistique et grammaticale de transformations d'arbres.

Dans WebContent

Réalisations Logicielles à venir

- Web service pour l'extraction. En cours de développement.
- Interface utilisateur pour l'interaction. Extension Firefox.