BIOTIM

Exploitation de Gisements Texte-Image en Biodiversité

http://www-rocq.inria.fr/imedia/biotim/

ACI Masses de données, appel 2003

Objectifs du projet

- Analyse et structuration de masses de textes et de masses d'images en biodiversité
- Acquisition de sur-couches sémantiques offrant la possibilité de créer des liens entre modalités
- Interrogation pluri-modale exploitant la complémentarité entre textes et images

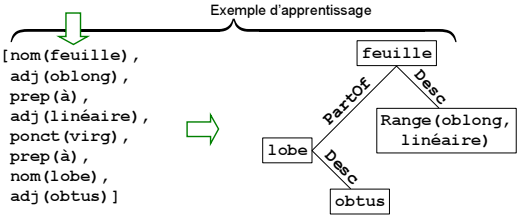
Analyse et structuration de masses de textes

- Textes analysés : extraits des volumes de *Flores*

Sections retenues (en vert) après structuration logique d'une page de Flores

- Phases préparatoires : structuration logique, chaîne morpho-syntaxique
- Étape d'extraction terminologique et d'identification de relations gouverneur-gouverné
- Analyse syntaxique : après 15 passages, la couverture passe de 36% à 68% sur 80000 phrases. Utilisation de la fouille d'erreurs, spécialisation de méta-grammaire, meilleures analyses pré-syntaxiques
- Extraction de classes sémantiques : s'appuie sur la similarité des contextes syntaxiques, sur les constructions d'intervalles (« *de* X *à* Y » ⇒ X et Y de même nature) et sur les marqueurs linguistiques (« *en forme de* X »). Amorçage : germes de classes. Fouille d'erreurs pour trouver les contextes les plus probables
- Processus général d'acquisition d'ontologie textuelle : texte pré-traité → règles de réécriture → arbres porteurs de l'information utile à la construction de l'ontologie
- Types de relations dans les structures arborescentes :
  1. Relation binaire `partie_de` (nom ↔ nom)
  2. Triplet `<nom, attribut, valeur>` (nom ↔ adjectif)
- Apprentissage règles de réécriture à partir d'exemples :

« feuilles oblongues à linéaires, à lobes obtus »



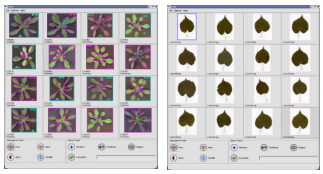
Analyse et structuration de masses d'images

- Méthodes adaptées de segmentation grossière et fine

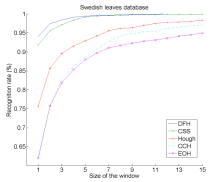
Grossière : séparation du fond pour décrire seulement la plante

Fine : pour mesures quantitatives (surface modifications couleur)

- Nouveau descripteur de forme DFH donnant de meilleurs résultats que d'autres descripteurs et ayant un coût réduit d'extraction et de comparaison



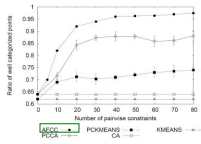
Recherche par similarité avec DFH sur 2 bases différentes ; l'image requête est en haut à gauche



Rappel avec différents descripteurs : DFH fournit les meilleurs résultats

- Nouvelle méthode de classification semi-supervisée active : meilleurs résumés avec moins d'interaction

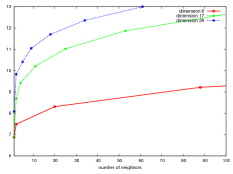
Représentants des 8 classes de la base de test (phénotypes de *Arabidopsis thaliana*)



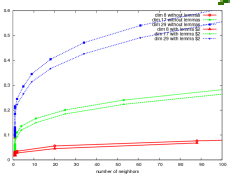
Comparatif : AFCC fournit les meilleurs résultats, avec le moins d'interaction

- Accélération de la recherche par le contenu image, en exploitant la multiplicité de points d'intérêt dans la requête

Exemple de points d'intérêt extraits



Gain E/S obtenu par requête multiple sur base réelle (10^6 points d'intérêt) et pour 3 dimensions (8, 17 et 29)



Temps CPU obtenu par requête multiple sur 10^6 points d'intérêt, sans et avec les améliorations proposés

Contacts nationaux et internationaux

- MNHN (Paris), CIRAD (Montpellier)
- Association ENDEMIA (Nouvelle Calédonie)
- *Nottingham Arabidopsis Stock Center* (Angleterre)
- Projet *Electronic Field Guide* (NSF, États-Unis)
- *Australian National Botanical Garden*

Publications scientifiques

- 2 publications internationales, 12 communications internationales, 4 communications nationales

Équipe Vertigo, CEDRIC (CNAM Paris)

UMR Génomique Végétale (INRA Evry)

IMEDIA et ATOLL (INRIA Rocquencourt)

IRD (Orléans)

Contraintes et Apprentissage, LIFO (Université d'Orléans)

Coordinateur : Michel Crucianu (Michel.Crucianu(à)inria.fr, Michel.Crucianu(à)cnam.fr)