

DOING: Données intelligentes

Porteur du sous-projet :

Mirian Halfeld Ferrari

MEMBRES (REGION)

- Anaïs Lefeuvre;
- Sylvie Billot (LIFO)
- Denis Maurel (LIFAT)
- Jean-Yves Antoine (LIFAT)
- Gaëtan Caillaut (LIFO)
- Guillaume Cleuziou (LIFO)
- Emmanuel Shang (LLL).
- Jacques Chabin (LIFO)
- Christel Vrain (LIFO)
- Agata Savary (LIFAT)
- Marcilio Pereira de Souto (LIFO)
- Anne Lyse Minard (LLL)

Autres: Joshua Amavi (ENTREPRISE PARIS)

Description du sous projet

Défi scientifique : Extraction d'Informations dans des textes médicaux pour l'acquisition, l'exploitation et maintenance d'une base de connaissances du domaine. Une mise en commun des compétences et spécialités régionales en Linguistique, TAL, Apprentissage et Bases de Données autour d'un défi commun.

Objectifs : Ce sous-réseau propose de rassembler des chercheurs de la région autour de la problématique des données intelligentes. Plus précisément, le sous-réseau s'intéresse aux étapes suivantes:

A- L'exploitation de corpus textuels médicaux dans le but de transformer des données en information. DOING propose de prendre en considération les aspects suivants:

- Obtention des métadonnées;
- Obtention d'une base ontologique;
- Obtention d'un schéma selon lequel les informations extraites des documents médicaux sont organisées dans une forme plus structurée (forme équivalente à un graphe d'attributs);

Cette étape fait appel notamment à des techniques d'extraction d'entités-nommées; d'acquisition (semi-)automatique de taxonomies lexicales; et de classification automatique.

B- L'exploitation des informations (métadonnées, données structurées, schéma...) obtenues par l'étape précédente (A) dans le but de transformer l'information en connaissance. Une base de connaissance devrait:

- assurer la qualité des informations;
- offrir des méthodes permettant l'interrogation et l'analyse des informations;
- offrir des mécanismes pour assurer l'évolution cohérente des connaissances.

DOING propose d'aborder ces trois aspects en faisant appel, notamment, à des techniques de requêtes, contraintes et mises à jour de bases données ainsi qu'aux langages formels et la logique. Le liens avec les techniques d'apprentissage sont à mettre en evidence par le réseau.

Aspects pluridisciplinaires:

Linguistique, TAL, Apprentissage et Bases de Données (voir objectifs)

Actions prévues (montage des projets, co-encadrements, colloques, etc.) :

- Ateliers de travail (sous-groupe ou tout le groupe)
- Co-encadrement des stages de recherche
- Montage des projets (nationaux ou internationaux)
- Préparation d'une proposition (sur la thématique du sous-réseau) à un "appel à workshop" lié à une conférence internationale reconnue.

Éléments à l'entrée du sous-projet (solutions, collaborations, résultats précédents, etc.) :

Les chercheurs en TAL (LIFAT et LIFO) travaillent déjà en collaboration avec des chercheurs en linguistique (LLL). Les chercheurs en apprentissage et bases de données du LIFO possèdent des résultats "mono-disciplinaires" sur le sujet; mais certains collègues d'apprentissage ont déjà collaboré dans le domaine du TAL. Les chercheurs impliqués dans ce sous-réseau souhaitent mettre en place plus d'interaction pour une montée en compétence régionale sur la problématique de données intelligentes.

Résultats attendus du sous-projet :

- Proposition d'une chaîne de traitement pour la transformation des données textuelles en connaissance: définitions des méthodes (formelles) et algorithmes.
- Application des méthodes et prototypes développés sur les données fournies par l'entreprise
- Connaissances approfondies des compétences et domaines de spécialité des autres chercheurs des domaines connexes et complémentaires → vers une réponse commune à des appels à projets futurs autour d'une offre scientifique élargie.
- Avancées scientifiques dans les domaines respectifs et émergence possible de nouvelles approches hybrides (ex. approches de fouille de données sur des Bases de Données « graphes »).

Collaborations nationales et internationales pouvant être en soutien au sous-projet :

- ETIS (Cergy-Pointoise): Dominique Laurent (Bases de données)
- GREYC (Caen) → Gaël Dias (Recherche d'Information)
- IRISA (Rennes) → Peggy Céliier (TAL et apprentissage); Laurent D'Orazio (Bases de données)
- IRIT (Toulouse) → José Moreno (Recherche d'Information)
- LIG (Grenoble): Genoveva Vargas-Solar (Bases de données)
- LIPN (Paris 13) → Davide Buscaldi (TAL et Extraction d'Informations) – Thierry Charnois (TAL et apprentissage)
- UGB (Université Gaston Berger, Saint-Louis, Senegal): Cheikh Ba (Bases des données)
- UFPR (Universidade Federal do Paraná, Brésil): Carmem Hara (Bases de données)
- UFRN (Universidade Federal do Rio Grande do Norte): Martin Musicante (Traitement et évolution des données)
- Équipe CA du LIFO: soutien scientifique; intérêts thématiques à certains moments

Liens avec des entreprises pouvant être en soutien au sous-projet : société Ennov (Ennov Labs) → Gestion Documentaire