

[UMR IATE](#) et [UMR ASB](#)

2, place Viala.

34 060 MONTPELLIER Cedex 02 - FRANCE

Sujet de stage de Master 2^{ème} année 2009-10
Association d'un indice de confiance aux tableaux de données
stockés dans un entrepôt de données ouvert sur le Web

L'UMR IATE (Ingénierie des Agro-polymères et Technologies émergentes) est une unité de recherche pluri-partenaires (INRA, CIRAD, Supagro, Univ. Montpellier 2) et pluridisciplinaire structurée selon cinq axes complémentaires qui étudient les grandes phases successives de la transformation des produits alimentaires et non alimentaires. Les travaux de recherche de l'UMR IATE sont donc au cœur des enjeux sociétaux et environnementaux actuels ayant pour objectifs le développement de procédés économisant l'énergie, l'élaboration de produits alimentaires de grande qualité sanitaire et nutritionnelle et d'emballages bio-dégradables. L'axe 5 « Représentation de connaissances et raisonnement » de l'UMR IATE développe des méthodes et des outils d'aide au pilotage global de filières de transformation.

Il travaille en étroite collaboration avec l'équipe MISAA (Méthodes Informatiques et Statistiques pour les Agrosystèmes et l'Agro-alimentaire) de l'UMR ASB (Analyse des Systèmes et Biométrie) qui développe des méthodes mathématiques, statistiques et informatiques pour l'analyse et l'aide à la décision des systèmes relevant de l'agronomie et de l'environnement, avec un accent particulier sur la modélisation, les systèmes dynamiques et les systèmes complexes.

Un des sujets de recherche commun de l'axe 5 et de l'équipe MISAA porte sur la conception et la réalisation d'entrepôts de données expérimentales ouvert sur le Web et s'articule autour des thèmes suivants :

- l'enrichissement semi-automatique d'un entrepôt par des données extraites du Web,
- la représentation de données imprécises en XML/RDF en utilisant le formalisme des sous-ensembles flous,
- l'interrogation flexible d'un entrepôt de données XML/RDF ouvert sur le Web.

L'axe 5 « Représentation de connaissances et raisonnement » a développé un certain nombre d'applications dans le cadre du projet ANR WebContent (<http://www.webcontent.fr/>), en partenariat avec l'unité INRA Mét@risk, qui seront utilisées dans ce travail.

L'ouverture sur le Web des entrepôts de données soulève la question de la confiance que l'on peut accorder aux sources de données issues du Web. Dans ce contexte, l'équipe « Représentation de connaissances et raisonnement » propose un stage dont l'objectif est de proposer une méthode permettant d'associer à des tableaux de données extraits du Web un indice de confiance. Cet indice de confiance permettra de décider si ces données sont suffisamment fiables pour être prises en compte dans des outils de simulation.

Les tableaux de données proviennent de documents de natures diverses (articles scientifiques soumis à comité de lecture, supports de cours, présentations scientifiques, ...) qui n'ont pas tous la même valeur scientifique. De plus, dans une publication donnée, des éléments contextuels plus ou moins structurés peuvent être pris en compte : facteur d'impact du journal dans lequel est publié l'article, nombre de fois où l'article est cité (hyperliens, section bibliographie), protocole expérimental, nombre de répétitions des expérimentations. On cherchera à déterminer plusieurs critères (statistiques, textuels, de dire d'experts) qui permettront d'associer un indice de confiance aux tableaux de données. La construction de l'indicateur de confiance sera réalisée en s'appuyant

sur la théorie des croyances. La théorie des fonctions de croyance initiée par Dempster (Dempster, 1967) et renforcée par Shafer (Shafer, 1976) et Smets (Smets et Kennes, 1994) propose un formalisme approprié pour manipuler l'incertitude associée aux critères. Elle permet en effet de définir de manière élaborée l'incertitude associée à un facteur de confiance, notamment via des outils tels que les fonctions de croyances conditionnelles (éventuellement modélisée via des réseaux causaux de fonctions de croyance) ou encore des méthodes d'affaiblissement (appauvrissement de l'information) très riches (Mercier et al., 2008).

Missions à effectuer

Le stage se déroulera en trois parties.

Premièrement, le/la stagiaire devra prendre connaissance de la méthode d'annotation sémantique de tableaux (Hignette et al. 2009) et d'interrogation flexible de ces tableaux annotés (Buche et al. 2009) dirigée par une ontologie de domaine, développée dans l'équipe. Il/elle devra également faire une étude bibliographique des travaux portant sur la théorie des fonctions de croyance. Deuxièmement, il/elle aura en charge la conception et la mise en œuvre d'une méthode de calcul d'un indice de confiance associé à un tableau de données du Web. Troisièmement, il/elle évaluera la pertinence de la méthode proposée en concevant et en mettant en œuvre une procédure de validation à partir de corpus de documents fournis par les partenaires de l'équipe.

Afin de réaliser ces travaux, l'étudiant aura également à interagir avec des experts des domaines d'application, dans le but d'identifier les critères implicites que ces derniers utilisent pour donner une confiance aux données qu'ils rencontrent et manipulent.

Profil et compétences souhaités

Etudiant en BAC+5, avec un parcours orienté informatique/mathématiques appliquées, attiré par les applications. Une bonne connaissance du langage R et/ou des théories de l'incertain constituerait un atout.

Bibliographie

- P. Buche, J. Dibia-Barthélemy, H. Chebil (2009). Flexible SPARQL querying of Web data tables driven by a domain ontology. FQAS'09 (Flexible querying and answering systems). To appear in LNCS.
- A. Dempster (1967). Upper and Lower probabilities induced by a multi-valued mapping. Ann. Math. Statist. Volume 38, Number 2 (1967), 325-339
- G. Hignette, P. Buche, J. Dibia-Barthélemy, O. Haemmerlé: Fuzzy Annotation of Web Data Tables Driven by a Domain Ontology. ESWC 2009: LNCS 5554 : 638-653
- D. Mercier, B. Quost and T. Denoeux (2008) Refined modeling of sensor reliability in the belief function framework using contextual discounting, Information Fusion, 9, 246–258
- G. Shafer (1976). A mathematical theory of evidence. Princeton University Press, New Jersey. 1976
- P. Smets and R. Kennes (1994). The transferable belief model. Artificial Intelligence. 66, 191-234.

Contacts :

Patrice Buche et Sébastien Destercke, UMR IATE axe 5 et Brigitte Charnomordic, UMR ASB

{buche, destercke, bch}@supagro.inra.fr