Description des sujets de recherche et du sujet de thèse

L'objectif de ma thèse consiste à étudier l'apport de l'apprentissage automatique, réalisé sous la supervision partielle d'un expert, afin de structurer les données complexes disponibles sur Internet, et plus particulièrement les documents comportant du texte et des images. Pour traiter cette question, l'approche que l'on envisage est d'utiliser des méthodes dites de regroupement conceptuel (conceptual clustering). Suivant cette approche, des travaux récents cherchent à structurer l'information disponible à partir de textes (livres, dictionnaires, annuaires, lexiques, etc.) sous la forme de hiérarchies de concepts afin, par exemple, d'améliorer l'indexation de ces documents et/ou de constituer des ontologies. Il a été clairement montré que ce processus ne peut être totalement réalisé de manière automatique et qu'un expert doit intervenir à un niveau ou à un autre de la chaîne de traitement. La piste originale que nous proposons pour la thèse est, d'une part, d'autoriser des recouvrements dans les concepts extraits des textes, qui serviront de « briques » à la hiérarchie construite, et, d'autre part, d'utiliser une supervision partielle sous la forme de codes (ou tags) qui sont donnés par l'expert non informaticien sur une partie des données en fonction de sa problématique de recherche.

Les différents parties qui composent la problématique de la thèse ont été abordées et étudiées en parallèle. Une des premières conclusions que nous avons tirée est que les connaissances a priori peuvent être utilisées à plusieurs niveaux pour améliorer les résultats des algorithmes d'apprentissage automatique. Parmi eux, les plus importantes : la création semi-supervisée de la représentation numérique des individus, l'injection des connaissances dans les algorithmes (algorithmes semi-supervisées), utilisation de la dimension temporelle comme une connaissance supervisé et la détection des évolutions typiques.

Création de la représentation numérique

Parmi les éléments du document complexes, on s'est intéressé à la façon d'introduire des connaissances expert pour deux des plus importants : le texte et l'image.

Le texte. La littérature propose une multitude d'algorithmes d'extraction/construction des thématique à partir de textes, mais elle ne propose pas de mesures pour les évaluer automatiquement. Nous avons proposé des mesures basées sur de l'information experte présentée sous forme de hiérarchies de concepts (ontologies, ex. WordNet). En se basant sur la distance entre les nœuds dans l'arbre des concepts, des scores peuvent être calculés automatiquement pour évaluer les thématique et leur pertinence. Ainsi, les mesures peuvent être utilisées comme étape de filtrage pour améliorer les algorithmes de construction des thématique.

La recherche sur la partie évaluation de thématique s'est matérialisée sous la forme de deux articles dans des conférences internationales **[d][e],** nationales **[f]** et un article de journal **[b]**. La recherche sur la détection des thématiques a été valorisée par la publication d'un chapitre de livre sur l'extraction des thématiques pour la construction automatique des ontologies **[e]**.

L'image. Pour traiter les images des documents complexes, il est nécessaire de les rendre dans un format compatible avec les données textuelles. Pour cela on a choisi d'utiliser une représentation « bag-of-features » (surnommé représentation avec des mots visuels). En utilisant un détecteur des régions affines, des points d'intérêt sont extraits à partir des images, puis ils sont décrits en utilisant un descripteur SIFT. Un vocabulaire visuel est créé à partir de la collection des images et il va servir comme une clef de transformation des images dans une description numérique.

Nous avons proposé deux algorithmes. Le premier introduit des connaissances expertes au niveau de la création du vocabulaire visuel. L'objectif de cet algorithme est d'augmenter le nombre de points d'intérêt relatifs aux objets pendant la construction de vocabulaire visuel. Pour chaque

objet on construit un vocabulaire dédié uniquement à partir des images contenant cet objet et à la fin on concatène les vocabulaires dédiés obtenus. Le second algorithme filtre les points d'intérêt qui appartiennent au fond de l'image afin d'obtenir une meilleure représentation numérique des objets. Cet algorithme utilise des méthodes inspirées de la reconnaissance d'objet dans des photos (et non le positionnement spatial) pour réduire le bruit et la dimensionalité de construction des vocabulaires visuels.

Les deux approches ont été proposées dans un article qui est en cours d'évaluation pour un journal international **[h]**.

Construction des attributs

Même si on a utilisé l'information expert au moment du passage à une représentation numérique, une qualité insuffisante vient souvent du fait que les attributs utilisés sont inadéquats pour représenter les individus en question. D'où le besoin de ré-écrire l'ensemble des attributs à partir d'un ensemble de données.

Nous proposons une méthode non-supervisée de construction des attributs, où chaque nouvel attribut est une forme conjonctive des anciens attributs ou de leur négation. Le but est de réduire les co-occurrences entres les attributs dans la description des individus et d'augmenter la séparabilité du nouvel espace de représentation. Notre algorithme se base sur un schéma général inductif supervisé et propose un nouveau schéma de recherche et un constructeur des attributs, ainsi qu'une mesure d'évaluation d'un ensemble des attributs. Nous proposons, également, une méthode basée sur des tests statistiques pour déterminer automatiquement la valeur des attributs.

L'algorithme de construction des attributs s'est matérialisée sous la forme d'un article dans un journal international **[a]**.

Détection des évolutions typiques

Souvent, les données issue du Web, mais surtout ceux du domaine des Sciences Humaines et Sociales (SHS) ont une dimension temporelle. Une collection des entités est étudié durant une période de temps et décrite en utilisant un certain nombre des descripteurs numériques (*e.g.* pays dans un jeu de données politiques, hommes politiques dans les forums sur le Web). Un des intérêts est de détecter et structurer les évolutions typiques des entités durant la période étudié.

Nous proposons une méthode non-supervisé, basé sur du clustering semi-supervisé, pour détecter les évolutions. Les informations temporelles servent comme information expert et nous proposons une nouvelle mesure de dissimilarité pour prendre en compte, à la fois, la dimension temporelle et descriptive des données.

L'algorithme de construction des attributs s'est matérialisée sous la forme d'un article dans une conférences internationales **[c]**.

Les perspectives

Les perspectives à court terme sont d'intégrer l'information spatio-temporelle dans la construction des attributs, afin de générer de nouveaux attributs capable de décrire l'information d'un façon plus pertinent.

Les perspectives à long terme sont d'adapter les contraintes utilisées dans la littérature relative au semi-supervisé au cas recouvrant et hiérarchique.

Publications

Journaux internationaux :

[a] Rizoiu M.-A., Velcin, J. & Lallich, S. *Unsupervised Feature Construction for Improving Data Representation and Semantics.* Journal of Intelligent Information Systems (2013)**,** accepté , à apparaître.

[b] Musat, C., Rizoiu M.-A., & Trausan-Matu, S. *An Intra and Inter-Topic Evaluation and Cleansing Method.* Romanian Journal of Human - Computer Interaction, vol. 3 (2010) p.81 – 96.

Conférences internationales :

[c] Rizoiu, M.-A., Velcin, J. & Lallich, S. *Structuring typical evolutions using Temporal-Driven Constrained Clustering.* In: 24th International Conference on Tools with Artificial Intelligence (ICTAI 2012), Athens, Greece, pages 610--617, IEEE, 2012. **Best Student Paper Award.**

[d] Musat, C., Velcin J., Trausan-Matu, S., & Rizoiu M.-A. *Improving Topic Evaluation Using Conceptual Knowledge.* In 22nd International Joint Conference On Artificial Intelligence (IJCAI 2011). Barcelona, Spain. July, 2011.

[e] Musat, C., Velcin J., Rizoiu M.-A., & Trausan-Matu, S. Improving Topic Models using Conceptual Data. In 19th International Symposium on Methodologies for Intelligent Systems (ISMIS 2011). Warsaw, Poland. June 2011.

Conférences nationales :

[f] Rizoiu, M.-A., Velcin, J., & Chauchat, J.-H. *Regrouper les données textuelles et nommer les groupes à l'aide des classes recouvrantes*. In 10éme conférence Extraction et Gestion des Connaissances (EGC 2010), Hammamet, Tunisie ; Vol. E-19, p. 561-572. janvier, 2010.

Chapitres de livre :

[g] Rizoiu, M.-A. & Velcin, J. Topic *Extraction for Ontology Learning.* Chapitre dans le livre « Ontology Learning and Knowledge Discovery Using the Web: Challenges and Recent Advances » (2011) p. 38-61.

Soumis :

[h] Rizoiu M.-A., Velcin, J. & Lallich, S. *Visual Vocabulary Construction for Image Classification in a Weakly Supervised Context.* International Journal of Artificial Intelligence Tools (2012).