Sujet de TER - Master 2 Data Mining - 2016-2017

1. Clustering de données textuelles et quantitatives (J.Jacques & J. Velcin)
L'objectif sera de développer un algorithme de clustering à base de modèle probabiliste pour données textuelles et quantitatives; Pour cela, un modèle de type LDA pour les données textuelles sera couplé à un modèle gaussien pour les données quantitatives, avec une hypothèse d'indépendance conditionnelle entre ces deux types de variables. Le travail consistera à écrire le modèle, l'algorithme d'estimation et programmer ce dernier sous R. Des tests seront ensuite réalisé sur données simulées et réelles.
Réf : cours model-based learning + article D. Blei, A. Ng, and M. Jordan (2003). Latent Dirichlet allocation. Journal of Machine Learning Research.

2. Prise en compte des données manquantes en co-clustering (J.Jacques)
La prise en compte des données manquantes est un enjeu primordial de toute méthode de data mining; Dans un contexte de co-clustering de données quantitatives, nous verrons comment il est possible de prendre en compte des données manquantes au sein même de l'algorithme, et ainsi de les imputer de façon intelligente. L'objectif consistera à étudier le modèle des blocs latents ainsi que son estimation, de l'implémenter et de le tester sur données réelles et simulées.
Réf : M. Nadif and G. Govaert, "Model-Based Co-clustering for Continuous Data," Machine Learning and Applications (ICMLA), 2010 Ninth International Conference on, Washington, DC, 2010, pp. 175-180.
Ce sujet pourrait être continué sous la forme d'un stage recherche au sein du laboratoire.

3. Build a Classifier to predict Malicious User Behavious with TensorFlow and Apache Spark (Y. Badr)
présenté en cours.

4. Data Lakes vs. Data Vaults (J. Darmont)
présenté en cours.

5. Méthodes de coclustering et applications à des données textuelles. (J. Ah-Pine)
L'objectif de ce TER est d'établir un état de l'art sur les méthodes de coclustering, d'implémenter au moins deux méthodes classiques et de les tester sur des tâches de fouille de texte. Il est attendu de l'étudiant (1) un rapport présentant l'état de l'art, les méthodes étudiées de façon plus approfondies ainsi que les résultats sur les tâches de fouille de texte et (2) le code source en R ou Python mettant en oeuvre les algorithmes et les expériences.

6. Functional Manifold Representations (J. Cugliari)
A partir de l'article Nonlinear Manifold Representations for Functional Data, D. Chen & H.-G. Müller (2012) les étudiants devront comprendre la
proposition des auteurs et mettre en place les simulations qu'ils sont conduit.
Ensuite, ils évalueront l'approche à partir d'un jeu de données fonctionnelles issu du milieu industriel.

7. Mise en place d'une structure de calcul distribuée orientée DM (J. Cugliari)
Les étudiants monteront un cluster type beowulf à partir des ordinateurs bon

marchés (e.g. rasperri pi).
Le travail couvre la structure matériel de calcul et la structure logiciel. Un algorithme de calcul distribuée orientée DM devrai pouvoir tourner de manière transparente à l'utilisateur (cf http://coen.boisestate.edu/ece/files/2013/05/Creating.a.Raspberry.Pi-Based.Beowulf.Cluster_v2.pdf)

8. Prévision en ligne par agrégation des prédicteurs (J. Cugliari)
Nombreux modèles existent pour prévoir des données temporelles, i.e. anticiper le comportement futur des données qui ne sont pas encore observées. Une alternative au choix d'un seul modèle, c'est de les faire prévoir de manière indépendante pour ensuite agréger les prévision.
Les étudiants devront étudier cette approche, évaluer différents variants sur des données simulées ainsi que sur des données réels du
marché financier français.

9. Analyse comparative de flux d'actualité (J. Velcin)
Dans le cadre du projet pluridisciplinaire "Evolutions du journalisme à l'ère du numérique" (JADN), nous proposons de réaliser une étude comparative de l'actualité proposée par un même média (en l'occurrence, le Huffington Post) mais dans différentes langues (anglais, français et brésilien). L'idée du TER est de caractériser chacun des flux à l'aide de différentes mesures vues en cours, des plus simples (comptage des mots, ou des expressions, les plus fréquents, nombre d'auteurs des articles) aux plus avancées (calculées sur des thématiques extraites avec LDA), puis de réaliser une première analyse comparative sur la base de ces mesures. La comparaison se fera en partenariat avec des chercheurs en sciences de l'information et de la communication, spécialisés dans l'étude des médias. La base de données comporte aujourd'hui plus de 40 000 articles et billets de blog extraits du Huffington Post depuis juin 2016.
Ce sujet pourrait être continué sous la forme d'un stage recherche au sein du laboratoire.