

Extraction des Connaissances dans les Données (ECD)

Mathieu Roche

Cours ECD

2007/2008

Organisation du cours

- Historique des modules ECD et FDA
- Organisation générale du module ECD (Cours/TP)
- Intervenants : Mathieu Roche, Chedy Raïssi, Jacques Chauché et Violaine Prince
- Modalités du contrôle des connaissances
- Plan du cours
- Planning

Plan du cours

- Introduction à l'ECD
- Un processus global de Fouille de textes (TPs)
- Recherche d'Informations et Langage Naturel (Généralités, représentation des textes pour la classification, utilisation de connaissances linguistiques, etc) (Projet),
- Techniques de base de Fouilles de Données (TPs).

→ **Finalité** : Projet de classification de textes (utiliser les algorithmes et représentations présentées en cours)

Calendrier

- Répartition des groupes de TPs et description du planning

Références

- Cours de Stéphane Tufféry : Data Mining and Scoring
- Cours de Jérôme Azé (Université de Paris-Sud).

La fouille de données aujourd'hui

La fouille de données ou l'exploitation du réel pour mieux le comprendre !

- De l'infiniment petit (génomique) à l'infiniment grand (astrophysique).
- Du plus quotidien (analyse de tickets de caisse) au plus confidentiel (pilotage automatique).
- Du plus ouvert (données d'internet) au plus confidentiel (données propres à une entreprise).

La fouille de données aujourd'hui

- **Spécificités de la fouille de données aujourd'hui :**
 - Grandes quantités de données disponibles (bases de données, textes, logs, traces diverses, etc.)
 - Etude de données complexes :
 - Thèmes spécialisés.
 - Types de données : structurées, semi-structurées, libres.
- **Conséquences :**
 - Demande d'interventions d'experts à tous les niveaux : **acquisition des données, choix des paramètres à utiliser** (nécessité de mettre en oeuvre des approches coopératives), **validation des résultats.**

Sondage : la fouille de données aujourd'hui (sources : <http://www.kdnuggets.com/>)

Scaling up DM algorithms for huge data (46)	41%
Mining text (33)	29%
Automating data cleaning (30)	27%
Dealing with unbalanced and cost-sensitive data (29)	26%
Mining data streams (20)	18%
Mining links and networks (19)	17%
Unified theory of DM (18)	16%
DM for biological problems (16)	14%
DM with privacy (10)	8.9%
Mining images (8)	7.1%
DM for security applications (6)	5.4%
Distributed (multi-agent) DM (4)	3.6%
Other (21)	8%

Sondage : les données traitées aujourd'hui (sources : <http://www.kdnuggets.com/>)

table data, fixed # of columns (103)	82%
time series (51)	40%
text, free-form (42)	33%
transactions (association rules) (38)	30%
web clickstream (21)	17%
spatial data (2-D, 3-D) (20)	16%
web content (19)	15%
email (16)	13%
XML data (16)	13%
links or networks (14)	11%
anonymized data (14)	11%
images/video (8)	6%
music and audio (8)	6%
other (21)	17%