UNIVERSITÉ DE SHERBROOKE

Département d'informatique

IFT501
Recherche d'information et forage de données

Plan de cours

Professeur : Shengrui Wang
Courriel : shengrui.wang@usherbrooke.ca
Local : D4-1018-1
Téléphone : (819) 821-8000 poste 62022
Consultation : mercredi 13h30-16h

Session : Automne 2016

Horaire : Lundi 14h30 à 16h20 Salle : D3-2033
Mardi 10h30 à 12h20 Salle : D3-2033

Description

Voir l'annuaire : http://www.usherbrooke.ca/programmes/cours/IFT/ift501.htm

Crédits : 3

Organisation Cours : 3(+1) heures/semaine
Travail personnel : 6 heures/semaine

Préalables : (IFT 339 et STT 389) ou (IFT 436 et STT 418)

1 Présentation

1.1 Mise en contexte

De nos jours, toutes les entreprises collectent des masses importantes de données. Ces mégas bases de données, qui ne cessent d'augmenter jour après jour, sont peu exploitées, alors qu'elles cachent des connaissances décisives face au marché et à la concurrence. Il est évident que la plupart des techniques d'analyse de données traditionnelles échouent à extraire de l'information pertinente sur ce type de données. Pour combler ce besoin, une nouvelle discipline émerge : le forage de données (Data Mining). Le forage de données est l'ensemble des algorithmes et méthodes destinés à l'exploration et l'analyse de grandes bases de données informatiques en vue de détecter dans ces données des règles, des associations, des tendances inconnues (non fixées *a priori*), des structures particulières restituant de façon concise l'essentiel de l'information utile.

1.2 Objectifs généraux

Ce cours vise à initier l'étudiante et l'étudiant aux concepts fondamentaux de forage de données et de recherche d'information et d'appliquer ces notions à des problèmes concrets.

1.3 Objectifs spécifiques

1. comprendre le processus d'extraction des connaissances dans les bases de données;
2. maîtriser les techniques d'analyse des données transactionnelles, plus particulièrement la recherche des règles d'association;
3. maîtriser des techniques de "clustering" et de classification les plus utilisées dans la pratique;
4. maîtriser des techniques de prétraitement des données et de réduction de la dimension, essentiellement la gestion de la redondance et la sélection des attributs;
5. comprendre la problématique des données de grande dimension et maîtriser des techniques de "clustering" utilisées sur ce type de données ;
6. se familiariser avec des algorithmes de classement des pages Web tel que HITS et PageRank ;
7. se familiariser avec des algorithmes pour les systèmes de recommandation.

1.4 Contenu détaillé

Le cours se concentre sur les concepts clés de forage de données et de la recherche d'information en les illustrant à travers des exemples et des descriptions simples des algorithmes les plus populaires.

Thème	Contenu	Heures	Objectifs
1	**Concepts de base** Processus d'extraction des connaissances dans les bases de données. Prétraitement des données. Mesures de similarités. Techniques de réduction de la dimension. Techniques de sélection d'attributs. Introduction de Weka (logiciel de forage de données)	4h (semaine 1 et 2 : 29 août – 5 sept.) **Pas de cours le 5 septembre (fête de travail)**	1, 4
2	**Analyse d'association** Introduction aux données transactionnelles. Principe de l'algorithme « Apriori » et l'algorithme « FP-growth ». Recherche et extraction des règles d'associations.	8h (semaines 3 et 4 : 12 et 19 sept.) + **Projet 1** dans la semaine du 12.	2
3	**Clustering (méthodes de base et méthodes avancées)** Clustering par partition et hiérarchique. Clustering basé sur la densité. Clustering des données catégoriques et transactionnelles. Clustering dans les sous-espaces : *Projected Clustering*. « Co-clustering ». Validation des résultats de clustering Détection des anomalies (outlier detection)	12 à 14h (sem. 5-8 : 26 sept, et 3 octobre) + **Projets 2** et **un devoir** ; Continuer après l'intra	3, 4, 5

		(sem. du 24 et 31 octobre)	
Intra	**Intra durant la semaine du 8-15 octobre (pas de cours)**	Date à déterminer	
4	**Systèmes de recommandation** Principes et éléments de bases. Principe des algorithmes basés sur le contenu. Approches basées sur le filtrage collaboratif : méthodes basées sur la ressemblance directe, méthodes basées sur la sémantique latente.	6h (sem. 8 et 9 : 7 et 14 nov.) + **Projet 3**	7, 4
5	**Classification supervisée** Arbres de décisions. *K* plus proche voisin.	4h (sem. 10 : 21 nov.)	3
6	**Web mining** Principes et éléments de bases. Algorithmes de classement des pages Web : HITS et PageRank. Extraction des communautés dans le Web.	*4h* (sem. 11 : 28 nov.)	6
Final	**Examen final le 6 décembre, durant le cours**		

2 Organisation

2.1 Méthode pédagogique

- Le cours sera donné sous forme d'exposés magistraux.
- Le cours comporte 3 travaux pratiques et un devoir. Le but principal de ces travaux est de consolider la compréhension des concepts vus en cours.
- Les mercredis après-midis sont réservés pour des périodes de consultation. Cependant, les étudiants sont les bienvenus pour consulter le professeur n'importe quand durant la semaine. **L'utilisation du courriel pour poser des questions sur le cours ou sur les travaux est fortement déconseillée !**

2.2 Calendrier du cours

SVP, voir le tableau des contenus détaillés pour le calendrier des prestations. L'examen périodique (intra) au lieu la semaine du 8 octobre. La semaine du 17 octobre est la semaine de relâche. **Il n'y aura pas de cours IFT501 le 8 septembre et le 12 décembre (et une autre journée à déterminer). L'examen final aura lieu le 6 décembre à 10h30.**

2.3 Évaluation

- Travaux pratiques et devoirs : 30%
- Examen périodique : 30%
- Examen final : 40%

2.4 Échéancier APPROXIMATIF des travaux

TP	Réception du problème : semaine de	Thème	Remise du travail (code + rapport) : semaine du	Pondération
1	12-09-2016	Analyse d'association	26-09-2016	6%
2	26-09-2016	Devoir		6%
3	26-09-2016	Clustering	31-10-2016	12%
4	7-11-2016	Système de recommandation	21-11-2016	6%

Directives particulières

- Les travaux pratiques doivent normalement être effectués seul ou par équipe de deux étudiant(e)s. L'équipe de trois personnes n'est permise que pour des circonstances exceptionnelles.
- Les travaux pratiques devraient normalement être programmés en Weka, ou en C ou C++ (l'environnement est à déterminer). Cependant, d'autres langages seraient admissibles. SVP parlez avec moi à l'avance !
- Le matériel à soumettre pour chaque TP inclut : le code et un rapport.
- La remise de chaque travail sera effectuée par le "turnin".

3 Documentation

1. Acétates du cours dans le répertoire public du DI.
2. P-N Tan, M. Steinbach, V. Kumar, Introduction to Data Mining, Addison Wesley, 2006.
   Ce livre contient plus que 50% de la matière du cours (thèmes : 1, 2, 3 et 4) et sera donc "obligatoire".
   Le livre est disponible au magasin Coop. Note que ce livre a été aussi utilisé pour le cours IFT603 (Techniques d'apprentissage) entre 2008 et 2011.
   Voir le site web du livre http://www-users.cs.umn.edu/~kumar/dmbook/index.php

2. J. Han, M. Kamber, Data mining: concepts and techniques, Morgan Kaufmann Publishers, 2006. *2nd edition.*

3. Ian H. Witten, Eibe Frank, Mark A. Hall, Data Mining: Practical Machine Learning Tools and Techniques (Third Edition), Morgan Kaufmann, January 2011, 629 pages, ISBN 978-0-12-374856-0.

Les articles suivants sont disponibles sur le web.

4. A.K. Jain, R.P.W. Duin and J. Mao, "Statistical Pattern Recognition: A Review", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, no. 1, pp. 4-37, 2000. (*thème 3 & 4*)

5. H. Liu and L. Yu, "Toward Integrating Feature Selection Algorithms for Classification and Clustering", IEEE Transactions on Knowledge and Data Engineering, vol.17, no4, pp. 491-502, 2005. *(thème: 5)*

6. Mohamed Bouguessa and Shengrui Wang, "Mining Projected Clusters in High Dimensional Spaces", IEEE Transactions on Knowledge and Data Engineering, 2008 *(thème: 5)*

7. Sergey Brin and Lawrence Page, "The Anatomy of a Large-Scale Hypertextual Web Search Engine", Computer Networks and ISDN Systems, vol. 30, no. 1-7, pp. 107-117, 1998. *(thème: 6)*

8. J. M. Kleinberg, "Authoritative Source in a Hyperlinked Environment", Journal of ACM, vol. 46, no. 5, pp. 604-632, 1999. *(thème: 6).*

9. Christian Desrosiers and George Karypis, "A Comprehensive Survey of Neighborhood-based Recommendation Methods", dans Recommender Systems Handbook, part 1, 107-144, 2011. (*thème: 7*)