UNIVERSITÉ DE SHERBROOKE

Département d'informatique

IFT 501 — Recherche d'information et forage de données

Plan de cours
Automne 2018

Enseignants :

Shengrui Wang
Courriel : shengrui.wang@usherbrooke.ca
Local : D4-1018-1
Téléphone : 819-821-8000 (62022)
Site du cours : Répertoire Public
Disponibilité : à déterminer

Responsable(s) : Shengrui Wang

Horaire :

Exposé magistral :	Mardi	8h30 à 10h20	Salle D3-2037
	Mercredi	10h30 à 12h20	Salle D3-2037

Description officielle de l'activité pédagogique[1]

Cibles de formation :	Comprendre et maîtriser les méthodes courantes pour la recherche d'information et la prospection de données.
Contenu :	Principes de la recherche d'information. Sélection des documents pertinents. Modèles booléen, vectoriel, probabiliste, logique. Évaluation des performances. Analyses linguistiques, syntaxiques et sémantiques. Moteurs de recherche. Processus de forage. Techniques de forage : caractérisation du forage descriptif, prétraitement de données, recherche et extraction des règles d'association, méthodes pour la classification et la prédiction, analyse de faisceau. Défis et outils du forage de données. Réalisation d'une application d'envergure.
Crédits	3
Organisation	3 heures d'exposé magistral par semaine 6 heures de travail personnel par semaine

Préalable(s) IFT 436 et STT 418

[1] https://www.usherbrooke.ca/admission/fiches-cours/IFT501/

1 Présentation

Cette section présente les objectifs spécifiques et le contenu détaillé de l'activité pédagogique. Cette section, non modifiable sans l'approbation d'un comité de programme du Département d'informatique, constitue la version officielle.

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 spécifiques

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.

À la fin de cette activité pédagogique, l'étudiante ou l'étudiant sera capable de :

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.

1.3 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	Travaux
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).	6	1, 4	Weka
2	**Analyse d'association et analyse des séquences** Introduction aux données transactionnelles ; Principe de l'algorithme « Apriori » et l'algorithme « FP-growth » ;	8	2	Analyse d'association

	Recherche et extraction des règles d'associations ; Extraction et utilisation des patrons significatifs pour l'analyse des séquences.			
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 des données de grandes dimensions et des données complexes ; Validation des résultats de clustering ; Détection des anomalies (outlier detection).	14	3,4,5	Clustering
Intra	**Durant la semaine, à partir de 6 octobre**	2		
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.	6	4,7	Système de recommandation
5	**Classification supervisée** Arbres de décisions ; *K* plus proche voisin.	4	3	
6	**Web mining** Principes et éléments de bases ; Algorithmes de classement des pages Web : HITS et PageRank ; Extraction des communautés dans le Web.	3	6	
Final	**Examen final durant le cours du 28 novembre (entre 10h30 et 12h30)**	**2**		

2 Organisation

2.1 Méthode pédagogique

- Le cours sera donné sous forme d'exposés magistraux.
- Le cours comporte trois travaux pratiques. Le but de ces travaux est de consolider la compréhension des concepts vus en cours.
- Les mardis 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

Le tableau suivant contient le calendrier des prestations. L'examen périodique (intra) a lieu entre les 6 et 13 octobre. L**'examen final aura lieu le mercredi 5 décembre à 10h30**.

Semaine	**Thème**	**Lecture**	**Dates de cours**	**Travaux**

1 et 2	1	Notes de cours (Concepts de base) Introduction de Weka (logiciel de forage de données).	Semaine du 27 août et du 3 sept.	TP0 : Weka
3 et 4	2	Notes de cours (Analyse d'association et analyse des séquences) -Partie 1 : Introduction, algorithme « Apriori » - Partie 2 : algorithme « FP-growth » - Partie 3 : patrons significatifs pour l'analyse des séquences.	Semaines du 3 au 17 sept.	TP1 : Analyse d'association
5 et 6	3	Notes de cours (Clustering) - Partie 1 : Clustering par partition, hiérarchique, basé sur la densité - Partie 2 : Clustering des données catégoriques et transactionnelles, et validation des résultats de clustering	Semaines du 17 sept. au 1 octobre	TP2 : Clustering
7	Intra	Durant la semaine à partir de 6 octobre	2h	
8		Semaine de relâche		
9 et 10	3	Notes de cours (Clustering) - Partie 3 : Détection des anomalies (outlier detection).; - Partie 4 : Clustering dans les sous-espaces : Projected Clustering ;	Semaines du 22 et du 29 octobre	
10 et 11	4	Notes de cours (Systèmes de recommandation)	Semaines du 29 oct. et du 5 nov.	TP3 : Système de recommandation
12	5	Notes de cours (Classification supervisée)	Semaine du 12 nov.	
13	6	Notes de cours (Web mining)	Semaine du 19 nov.	
14	**Final**	**Examen final durant le cours du 5 décembre (entre 10h30 et 12h30)**	**2 heures : 5 déc.**	

2.3 Évaluation

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

Conformément au règlement facultaire d'évaluation des apprentissages,[2] l'enseignant peut retourner à l'étudiante ou à l'étudiant tout travail non conforme aux exigences quant à la qualité de la langue et aux normes de présentation.

2 https://www.usherbrooke.ca/sciences/fileadmin/sites/sciences/Etudiants_actuels/Informations_academiques_et_reglements/2017-10-27_Reglement_facultaire_-_evaluation_des_apprentissages.pdf

Le plagiat consiste à utiliser des résultats obtenus par d'autres personnes afin de les faire passer pour sien et dans le dessein de tromper l'enseignant. Si une preuve de plagiat est attestée, elle sera traitée en conformité, entre autres, avec l'article 9.4.1 du Règlement des études[3] de l'Université de Sherbrooke. L'étudiant ou l'étudiante peut s'exposer à de graves sanctions dont automatiquement un zéro (0) au devoir ou à l'examen en question.

Ceci n'indique pas que vous n'ayez pas le droit de coopérer entre deux équipes tant que la rédaction finale des documents et la création du programme reste le fait de votre équipe. En cas de doute de plagiat, l'enseignant peut demander à l'équipe d'expliquer les notions ou le fonctionnement du code qu'il considère comme étant plagié. En cas de doute, ne pas hésiter à demander conseil et assistance à l'enseignant afin d'éviter toute situation délicate par la suite.

2.4 Échéancier des travaux

TP	Réception du problème : semaine de	Thème	Remise du travail (code + rapport) : semaine du	Pondération
1	10-09-2018	Analyse d'association	24-09-2018	10%
2	24-09-2018	Clustering	22-10-2018	10%
4	5-11-2018	Système de recommandation	19-11-2018	10%

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 Web".

2.5 Utilisation d'appareils électroniques et du courriel

Selon le règlement complémentaire des études, section 4.2.3[4], l'utilisation d'ordinateurs, de cellulaires ou de tablettes pendant une prestation est interdite sauf si leur usage est explicitement permis dans le plan de cours.

Comme indiqué dans le règlement universitaire des études, section 4.2.3[5], toute utilisation d'appareils de captation de la voix ou de l'image exige la permission du professeur.

Note : L'utilisation du courrier électronique n'est pas recommandée pour poser vos questions.

3 Matériel pour le cours

Aucun manuel n'est obligatoire. Toutes diapositives présentées dans le cours sont disponibles dans le répertoire public Public/Cours dont l'accès est décrit dans la page Web

http://www.usherbrooke.ca/informatique/intranet/ressources-et-documentation/faq/acces-aux-lecteurs-reseaux/

3 https://www.usherbrooke.ca/registraire/droits-et-responsabilites/reglement-des-etudes/

4 https://www.usherbrooke.ca/sciences/fileadmin/sites/sciences/documents/Intranet/Informations_academiques/Sciences_Reglement_complementaire_2017-05-09.pdf

5 https://www.usherbrooke.ca/registraire/droits-et-responsabilites/reglement-des-etudes/

4 Documentation et références

1- **Acétates du cours et des documents supplémentaires dans le répertoire public du DI**.
2- P-N Tan, M. Steinbach, V. Kumar, Introduction to Data Mining (Second Edition), Addison Wesley, 2018. Ce livre contient plus que 65% de la matière du cours (thèmes : 1, 2, 3 et 4). Voir le site web du livre http://www-users.cs.umn.edu/~kumar/dmbook/index.php.
3- J. Han, M. Kamber, Data mining: concepts and techniques, Morgan Kaufmann Publishers, 2006. 2nd edition.
4- 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.

5- 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).
6- 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)
7- Mohamed Bouguessa and Shengrui Wang, "Mining Projected Clusters in High Dimensional Spaces", IEEE Transactions on Knowledge and Data Engineering, 2008 (thème: 5)
8- 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)
9- J. M. Kleinberg, "Authoritative Source in a Hyperlinked Environment", Journal of ACM, vol. 46, no. 5, pp. 604-632, 1999. (thème: 6).
10- Christian Desrosiers and George Karypis, "A Comprehensive Survey of Neighborhood-based Recommendation Methods", dans Recommender Systems Handbook, part 1, 107-144, 2011. (thème: 7)

UNIVERSITÉ DE SHERBROOKE

L'intégrité intellectuelle passe, notamment, par la reconnaissance des sources utilisées. À l'Université de Sherbrooke, on y veille!

Extrait du *Règlement des études* (Règlement 2575-009)

9.4.1 DÉLITS RELATIFS AUX ÉTUDES

Un délit relatif aux études désigne tout acte trompeur ou toute tentative de commettre un tel acte, quant au rendement scolaire ou une exigence relative à une activité pédagogique, à un programme ou à un parcours libre.

Sont notamment considérés comme un délit relatif aux études les faits suivants :

a) commettre un plagiat, soit faire passer ou tenter de faire passer pour sien, dans une production évaluée, le travail d'une autre personne ou des passages ou des idées tirés de l'œuvre d'autrui (ce qui inclut notamment le fait de ne pas indiquer la source d'une production, d'un passage ou d'une idée tirée de l'œuvre d'autrui);

b) commettre un autoplagiat, soit soumettre, sans autorisation préalable, une même production, en tout ou en partie, à plus d'une activité pédagogique ou dans une même activité pédagogique (notamment en cas de reprise);

c) usurper l'identité d'une autre personne ou procéder à une substitution de personne lors d'une production évaluée ou de toute autre prestation obligatoire;

d) fournir ou obtenir toute aide non autorisée, qu'elle soit collective ou individuelle, pour une production faisant l'objet d'une évaluation;

e) obtenir par vol ou toute autre manœuvre frauduleuse, posséder ou utiliser du matériel de toute forme (incluant le numérique) non autorisé avant ou pendant une production faisant l'objet d'une évaluation;

f) copier, contrefaire ou falsifier un document pour l'évaluation d'une activité pédagogique;

[…]

Par plagiat, on entend notamment :

- Copier intégralement une phrase ou un passage d'un livre, d'un article de journal ou de revue, d'une page Web ou de tout autre document en omettant d'en mentionner la source ou de le mettre entre guillemets;
- reproduire des présentations, des dessins, des photographies, des graphiques, des données… sans en préciser la provenance et, dans certains cas, sans en avoir obtenu la permission de reproduire;
- utiliser, en tout ou en partie, du matériel sonore, graphique ou visuel, des pages Internet, du code de programme informatique ou des éléments de logiciel, des données ou résultats d'expérimentation ou toute autre information en provenance d'autrui en le faisant passer pour sien ou sans en citer les sources;
- résumer ou paraphraser l'idée d'un auteur sans en indiquer la source;
- traduire en partie ou en totalité un texte en omettant d'en mentionner la source ou de le mettre entre guillemets ;
- utiliser le travail d'un autre et le présenter comme sien (et ce, même si cette personne a donné son accord);
- acheter un travail sur le Web ou ailleurs et le faire passer pour sien;
- utiliser sans autorisation le même travail pour deux activités différentes (autoplagiat).

Autrement dit : mentionnez vos sources

Document informatif V.3 (août 2017)