INF7710 – Théorie et applications de la fouille d'associations

Plan de cours – Hiver 2011

GROUPE 40 VALTCHEV, Petko valtchev.petko@uqam.ca (514) 987-3000 1919 PK-4415

Jeudi, de 13h30 à 16h30

DESCRIPTION

La découverte d'associations est un aspect fondamental de la fouille de données. Ce cours met l'accent sur les bases théoriques de l'approche et sur les liens avec des problématiques de la théorie de la normalisation en bases de données, l'analyse formelle de concepts et les fonctions Booléennes. Problème générique de découverte d'associations et de la fouille de motifs fréquents. Variantes : motifs fermés, motifs maximaux, motifs clés ou générateurs. Structures algébriques mises en jeux : treillis Booléen, classes d'équivalence, correspondances de Galois, treillis de concepts, contextes. Approches de fouille de motifs : algorithmes par niveaux, algorithmnes verticaux, algorithmes hybrides. Représentations compactes pour les associations : base canonique, bases génériques et informatives. Famille réduites de motifs : motifs indérivables, motifs delta-libres, motifs sans disjonction, motifs k-libres. Applications de la fouille d'associations.

OBJECTIF

Ce cours vise à approfondir les connaissances de l'étudiant sur un domaine en pleine expansion qu'est la fouille de données. Le cours se focalise sur la découverte d'associations et de motifs fréquents qui est une discipline fondamentale de la fouille de données. L'accent est mis sur la présentation des diverses instanciations du problème général de la fouille ainsi que sur les fondements théoriques de l'approche et leurs liens avec des problématiques de la théorie de la normalisation en bases de données et l'analyse formelle de concepts.

Le but du cours est de permettre à l'étudiant de se familiariser avec la fouille de données en général à travers l'étude plus approfondie d'une des disciplines pertinentes, soit la fouille d'associations. Les objectifs concrets du cours peuvent être énoncés comme suit :

- ❏ d'approfondir la maîtrise de certains concepts fondamentaux en bases de données;
- ❏ de familiariser l'étudiant avec la démarche générale en fouille d'associations;
- ❏ de lui fournir des connaissances exploitables en conception de méthodes de fouille;
- ❏ de faire connaître à l'étudiant les plus récents développements dans le domaine;
- ❏ de permettre l'approfondissement d'un des thèmes de recherche dans le domaine;
- ❏ d'initier l'étudiant à la recherche à travers la rédaction d'un rapport sur un sujet d'actualité.

ÉVALUATION

Description sommaire	Date	Pondération
Résumé d'articles de recherche 1	Semaine 4	15%
Résumé d'articles de recherche 2	Semaine 6	15%
Travail de session : 1ère partie – Proposition de sujet	Semaine 7	10%
Travail de session : 2e partie – Présentation orale	Semaines 13 et 14	25%
Travail de session : 3e partie – Mémoire sur le sujet choisi	Semaine 15	35%

Le travail de session est réalisé par groupe de deux étudiants. Une liste de sujets potentiels et des recommandations seront fournis pendant la session. La qualité du français constitue un critère d'évaluation (pour un maximum de 10%). En cas de retard dans la remise des travaux, une pénalité de 5% par jour ouvrable sera appliquée. Un retard de plus d'une semaine ne sera pas accepté.

Les règlements concernant le plagiat seront strictement appliqués. Pour plus de renseignements, veuillez consulter les sites suivants : **http://www.sciences.uqam.ca/decanat/note_integrite.doc**

et **http://www.bibliotheques.uqam.ca/recherche/plagiat/index.html**

CONTENU

Ce cours est destiné aux étudiants ayant suivi au préalable un ou des cours de bases de données au niveau baccalauréat et, de préférence, un cours d'introduction à l'intelligence artificielle. Les sujets abordés sont :

- ❏ Introduction à la problématique de la fouille de données
- ❏ Principales tâches de la fouille et solutions de principe
- ❏ Problème générique de la fouille d'associations et ses défis techniques
- ❏ Approche de référence pour l'extraction d'associations
- ❏ Approches alternatives et familles réduites de motifs et d'associations
- ❏ Ordres partiels et treillis et leur rôle en fouille de motifs
- ❏ Analyse de concepts comme cadre fondamental pour fouille de motifs
- ❏ Associations, implications, dépendances fonctionnelles
- ❏ Motifs et associations structurés : séquences, arbres, graphes

- ❏ Fouille en présence de connaissances du domaine : les motifs généralisés
- ❏ Principales applications des motifs et associations : médecine, recherche d'information, recommandation, modélisation de l'utilisateur.

RÉFÉRENCES

V C TAN, P.-N., STEINBACH, M., et KUMAR, V. – *Introduction to Data Mining* – **Pearson (2005)** – **http://www-users.cs.umn.edu/~kumar/dmbook/**

V C GODIN, R. – *Systèmes de gestion de bases de données par l'exemple* – **Loze-Dion (2006)** – **http://www.info2.uqam.ca/~godin/livreEd2.html**

V C HAN J. et KAMBER, M. – *Data Mining : Concepts and Techniques, 2nd éd.* – **Morgan Kaufmann (2006).**

V C BERRY, M. et LINOFF, G. – *Mastering Data Mining* – **John Wiley & Sons (2000).**

V C HAND, D., MANNILA, H. et SMYTH, P. – *Principles of Data Mining* – **MIT Press (2000).**

V C CARPINETO, C. et ROMANO, G. – – *Concept Data Analysis : Theory and Applications* – **Wiley (2004).**

V C GANTER, B. et WILLE R. – – *Formal Concept Analysis, Mathematical Foundations.* – **Springer-Verlag (1999).**

A C FAYYAD, U., PIATETSKY-SHAPIRO, G. et SMYTH, P. – – *From Data Mining to Knowledge Discovery in Databases* – **AI Magazine 17(3): 37-54, 1996.**

A C AGRAWAL, R., IMIELINSKI, T. et SWAMI, A. – *Mining association rules between sets of items in large databases* – **Proceedings of the ACM SIGMOD International Conference on the Management of Data, Washington (DC), USA, pages 207–216, 1993.**

A C AGRAWAL, R., MANNILA, H., SRIKANT, R., TOIVONEN, H. et VERKAMO, A. – *Fast Discovery of Association Rules* – **U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, éds, Advances in Knowledge Discovery and Data Mining, pages 307–328. AAAI Press, Menlo Park (CA), USA, 1996.**

A C D'autres références (selon les sujets abordés) seront remises durant la session.

A : article – C : comptes rendus – L : logiciel – N : notes – R : revue – S : standard – U : uri – V : volume

C : complémentaire – O : obligatoire – R : recommandé