Fouille de données

Mots-clés

Bases de données, entrepôts de données, motifs fréquents, règles d'associations, réseaux sociaux

Description

L'objectif de cette activité est de proposer de nouvelles techniques de fouille de données ou de mesures d'intérêt pour les techniques existantes. Les données que nous traitons sont différentes, partant des bases de données classiques, jusqu'aux données complèxes fournies par les réseaux sociaux.

Dans ce contexte, nous nous intéressons particulièrement aux sujets suivants :

1. L'étude de motifs fréquents et des règles d'associations dans différents contextes :
    - règles d'associations construites à partir de requêtes fréquentes conjonctives dans les entrepôts de données ;
    - règles d'association disjonctives pour l'extraction de connaissances non fréquentes : dans la plupart des processus de fouille de règles d'associations, on s'intéresse aux motifs fréquents. Cependant, des motifs, qui individuellement ne sont pas fréquents, peuvent constituer un ensemble fréquent s'ils présentent des similitudes pertinentes. Dans cette étude, on s'intéresse à la fouille de règles d'associations disjonctives basées sur ce type de motifs non-fréquents.
2. La fouille de graphes pour les algorithmes de recommandation dans les réseaux sociaux : les réseaux sociaux jouent un rôle de plus en plus important dans les communications entre utilisateurs. Parmi les nombreuses applications de ce domaine, la recommandation d'un produit donné P à des utilisateurs d'un réseau social constitue un problème important et non trivial. En effet, ces réseaux étant de taille très importante, il n'est pas envisageable de parcourir le graphe entier pour déterminer les utilisateurs potentiellement intéressés par le produit P. Afin d'optimiser ce parcours, nous considérons dans cette étude qu'une ontologie est définie sur l'ensemble des produits et, à partir du « profil » du produit P, comparé aux profils des produits achetés par les différents utilisateurs, nous définissons des heuristiques de parcours de graphe qui évitent un parcours exhaustif tout en garantissant une bonne couverture de l'ensemble des utilisateurs potentiellement intéressés par le produit P.

Participants

Dominique Laurent (PU), Dimitris Kotzinos (PU), Claudia Marinica (MCF), Tao-Yuan Jen (MCF), Boris Borzic (IR), Jean-Philippe Attal (doctorant), Inès Hilali (doctorante)

Projets en cours

PARCOURS (2013-2015)

Mots-clés : web sémantique, ontologie, langages et outils du web sémantique, patrimoine, HADOC, CIDOC CRM

Description : PARCOURS (PAtrimoine culturel et Restauration-Conservation : Ontologie pour l'Usage d'un Référentiel commun aux différentes Sources de données) est soutenu par le LabEx PATRIMA qui vise à rassembler sciences de l'homme et sciences exactes au sein de la Fondation des sciences du patrimoine en structurant un grand nombre d'équipes autour de ces intérêts.
Actuellement, les bases de données et mesures scientifiques afférentes aux œuvres ou monuments sont souvent une juxtaposition de données, isolées les unes des autres et l'utilisateur a beaucoup de difficultés à mettre en relation ces travaux. A travers une approche ontologique, l'objectif principal de PARCOURS est de fournir un point de référence commun pour des sources d'information du patrimoine divergentes et incompatibles qui peuvent être ainsi comparées, harmonisées et interopérables. Ce système d'information doit permettre une interrogation unifiée des sources sur les métadonnées et sur la similarité de contenu simultanément ou non.

Participants : LRMH (porteur du projet), C2RMF, CRCC

Dates : 2013-2015

STIC Asie GOD - Indonésie, Malaisie, Vietnam, Univ. Clermont Ferrand, Univ. Paris Sud (2013-2015)

Mots-clés : Cloud Computing, optimisation de requêtes, entrepôts de données, enseignement à distance, données du patrimoine, applications médicales

Description : L'objectif du projet est d'étudier les techniques de cloud computing dans le cadre de l'optimisation de requêtes en utilisant un cache afin de stocker certaines réponses. Ce problème n'est pas nouveau, mais des résultats récents dans le cadre des entrepôts de données (en liaison très étroite avec les travaux sur les requêtes fréquentes) pour des requêtes SQL contenant des calculs d'agrégats permettent d'envisager de tirer profit des architectures étudiées dans le cadre des nuages pour utiliser et optimiser le contenu du cache de manière plus optimale que dans les approches standard proposées jusqu'alors. Les résultats théoriques seront validés dans différents entrepôts de données dédiés à l'enseignement à distance (Malaisie, Indonésie), les données du patrimoine (ETIS, LRI, Malaisie), ou les applications médicales (LIMOS, Vietnam).

Participants France : LIMOS-Univ. Clermond Ferrand (porteur du projet), LRI-Univ. Paris-Sud, ETIS-Univ. Cergy Pontoise

Participants Asie : HELP University (Kuala Lumpur, Malaisie), Univ. of Indonesia (Depok, Indonésie), Can Tho University (Can Tho, Vietnam)

Dates : 2013-2015

Thèses en cours

Ines Hilali (2010 -)

Dalia Sulieman (2009 -)

Collaborations

- LRI (Univ. Paris Sud),
- LIMOS (Univ. Clermont Ferrand),
- LINA (Univ. de Nantes),

- Tunisie (?),
- EISTI (?)