GDR I3 - Groupe 3.4
« Fouille de données »

12 juin 2003 – Paris

Jean-Marc Petit
Pascal Poncelet

Université Paris 5

Plan

- Fouille de données
  - Problèmes, Challenges
  - Un exemple : le passage à l'échelle
- Fonctionnement du GT
  - Aspects financiers
  - Aspects scientifiques
- Objectif du GT
- Objectif de cette réunion

Fouille de données

- Domaine en plein essor
  - Déluge de données : les données s'accumulent …
  - Pénurie de connaissance sur les données
- Nombreux domaines d'application
  - Biologie, Finance, Astronomie, « Panier de la ménagère » …
- De plus en plus de conférences et revues
- A l'intersection de plusieurs domaines, parmi lesquels
  - Apprentissage
  - Bases de Données
  - Statistiques

Fouille de données (cont.)

- Les données à traiter sont
  - de plus en plus volumineuses (e.g. nombre de transactions pour les règles d'association)
  - et complexe (e.g. nombre de variables ou de dimensions)

- Pose de nouveaux challenges, par exemple le passage à l'échelle d'algorithmes classiques

Passage à l'échelle des algorithmes

- Obtenir un comportement quasi-linéaire dans la taille des données
- Gestion mémoire
  - Mémoire centrale vs mémoire secondaire
  - Transfert d'un bloc <=> 10^6 instructions
- Trois grandes approches :
  - Algorithmique en mémoire externe / fichiers
  - Couplage faible BD/LP (SQL + LP)
  - Couplage fort BD/LP

Un exemple sur les tris

- Complexité en nombre d'opérations en O(n log(n))
- Que se passe-t-il quand n devient très grand ?
  - « Core dump » ☹
- Algorithmes de tri à plusieurs phases
  - Limites les coûts d'E/S
  - Gestion explicite de la mémoire
  - Opérations de base des SGBDs (couplage fort)

=> un exemple réussi de passage à l'échelle

Fonctionnement : aspects financiers

- Financement du groupe
  - Au prorata du budget du GDR I3 …
- Environ 2 journées thématiques par an
  - juin et novembre
- Remboursement en priorité pour les orateurs
  - Priorité affichée en faveur des jeunes chercheurs

Fonctionnement : aspects scientifiques

- Constitution d'un comité de pilotage
  - Apporte une caution scientifique au groupe
  - Valide les choix scientifiques
  - Membres : H. Briand, G. Hébrail, P. Gallinari, L. Lakhal , D. Laurent,E.M Nguifo, M. Sebag, D. Zighed
- Favoriser l'organisation de journées scientifiques
  - Orateurs invités
  - Appel à des exposés
    - processus ouvert via les listes de diffusion

Objectifs du GT

- Participer à l'animation de la communauté française en fouille de données
  - Doit être complémentaire
    - des conférences
    - des actions spécifiques, RTP, autres groupes
- A court terme
  - 2 journées scientifiques par an
  - 1 site Web : http://www.lgi2p.ema.fr/~poncelet/GDRI3FD
- A plus long terme
  - Organisation d'écoles d'été
  - Rédaction d'un ouvrage collectif

Thématiques du GT

- Extraction des connaissances
- Fondements théoriques de la fouille de données
- Problèmes du passage à l'échelle
- Applications de la fouille de données
- Pré-traitement et post-traitement des connaissances
- Gestion dynamique des connaissances
- …

Objectif de cette journée

- Faire le point des différentes actions entreprises en France autour de la fouille de données
- Présenter des thématiques de recherche
- Positionner ce GT dans le paysage français
  - RTP, AS du CNRS
  - EGC, AFIA, autre groupe du GDR I3
  - Autres ?

Programme

9h - 9h-30	*Accueil des participants*
9h30 - 10h	Présentation des Objectifs du Groupe
10h - 10h30	**Michèle Sebag** (LRI, Orsay) Leçons tirées de l'Action Spécifique Fouille de Bases de Données
10h30 - 11h	**Djamel Zighed** (ERIC, Lyon 2) Fouille de Données Complexes
Pause	
11h15-11h45	**Patrick Gallinari** (LIP6, Paris) Propositions autour des RTP
11h45-12h15	**Georges Hebrail** (ENST, Paris) Quelques problématiques de gestion des données pour la fouille des données

Programme (fin)

14h-14h30	**Engelbert Mephu NGUIFO** (CRIL, Lens) Fouille de Données et Treillis de Galois
14h30-15h	**Lotfi Lakhal** (LIF, Marseille) Bases de données et Fouille de données
15h-15h30	**Fabrice Guillet** (IRIN, Nantes) Evaluation de la Qualité des Connaissances par l'Intensité d'Implication
15h30-15h45	Pause
15h45-16h45	Discussions et Bilan de la Journée