TANAGRA : une plate-forme d'expérimentation pour la fouille de données

Ricco RAKOTOMALALA
Laboratoire ERIC
Université Lumière Lyon 2
5, av. Mendés France
69676 BRON cedex
e-mail : rakotoma@univ-lyon2.fr

Résumé

TANAGRA est un logiciel « open source » gratuit dédié à la fouille de données. Il s'adresse à deux types de publics. D'un côté, il présente une interface graphique aux normes des logiciels de fouille de données actuels, y compris les logiciels commerciaux, le rendant ainsi accessible à une utilisation de type « chargé d'études » sur des données réelles. De l'autre, du fait que le code source est librement disponible et l'architecture interne très simplifiée, il se prête à une utilisation de chercheurs qui veulent avant tout expérimenter de nouvelles techniques en améliorant celles déjà implémentées ou en introduisant de nouvelles. TANAGRA est opérationnel ; les versions stables sont disponibles sur le Web depuis janvier 2004 (http://eric.univ-lyon2.fr/~ricco/tanagra/). Ce site compte en moyenne une vingtaine de visiteurs par jour.

Mots-clés : TANAGRA, Logiciel « open source », Fouille de données, Expérimentation

Abstract

TANAGRA is a data mining software for practitioners and for researchers. It answers two specifications: a software with user friendly GUI for realistic studies; an "open source" software for researcher experimentations. Internal architecture is very simplified; users can easily add new features or data mining methods. TANAGRA is operative; stable versions are available on the web since January 2004 (http://eric.univ-lyon2.fr/~ricco/tanagra/). Website has about 20 visitors a day.

Keywords: TANAGRA, Open source software, Data mining, Experimentation

1 Introduction

La Fouille de Données, ou de manière générique l'Extraction de Connaissance à partir de Données (en anglais *Data Mining and Knowledge Discovery in Databases*), est un domaine de recherche qui a véritablement pris son essor au milieu des années 90. S'il est toujours possible de discuter quant à sa véritable originalité par rapport au traitement statistique des données qui existe déjà depuis longtemps (Hand et al., 2001), il est indéniable en revanche que son avènement s'est accompagné d'une forte accélération de la diffusion de logiciels spécialisés estampillés Data Mining. Les raisons sont multiples ; on pourra citer entre autres : la rencontre entre des communautés différentes (apprentissage automatique, bases de données, analyse de données, statistiques) ; le développement d'Internet qui a permis la diffusion à peu de frais des logiciels, avec pour certains des codes sources ; l'élargissement du champ d'application du traitement des données