Rapport du TER GMIN401 :

Intégration et optimisation d'algorithmes de classifications supervisées pour Weka

Par : ALIJATE Mehdi - NEGROS Hadrien - TURKI Batoul

31 Janvier 2014

Table des matières

Résumé

Ce sujet vise à intégrer et à optimiser des algorithmes de classifications supervisées de documents dans la suite logiciel WEKA. Ces algorithmes sont issus de travaux de recherche menés récemment au sein du LIRMM.

1 Introduction

La classification de documents est couramment utilisée afin de classer automatiquement des ressources en provenance d'un corpus.
Plusieurs formes de classification existent (par genre, par opinion, par thème...etc), et se font via des algorithmes de classifications spécifiques. Ceux-ci se basent sur des méthodes principalement numériques (probabilistes), avec des algorithmes de type mathématique ou basés sur la recherche d'information.
Ce TER vise justement à intégrer des algorithmes de classifications supervisées de documents dans la suite logiciel WEKA [1], se basant sur un nouveau modèle de classification à partir d'un faible nombre de document, intégrant de nouvelles pondérations adaptées.
Tout d'abord, il faudra explorer ce logiciel WEKA, pour une meilleure prise en main du code source, la maniabilité des classes et explorer une méthode d'ajout d'un algorithme dans l'API. Ensuite, nous nous pencherons sur le développement des différentes classes en établissant une méthodologie concrétisant le travail mené au laboratoire du LIRMM, s'en suivera une phase d'intégration et différents tests.

Ce présent mini-rapport présente un compte rendu de la première phase de notre travail, qui s'est déroulée entre notre dernière réunion le 24/01/14 et aujourd'hui.

2 Exploration de WEKA

Après la réunion du 24/01/14, nous avons établi un plan de travail pour bien mener et répartir les tâches de ce TER. Il a été décidé de le divisé en trois grandes parties succéssives et indissociabe. La première, qui est décrite ci-dessous consiste à explorer et prendre en main le logiciel WEKA, afin de pouvoir à la fin (le But) y rajouter les algorithmes qu'on aura développer lors de la deuxième partie, et qui seront tester et intégrer lors de la troisième.

2.1 L'API Weka/Sources avec Eclipse

//TODO

2.2 L'utilisation des classes

Une fois familiarisés avec l'API Weka, on a creusé un peu plus du côté des classes qui pourraient nous être utiles pour ce TER. Il s'agit des certaines classes présentes dans le package "weka.classifiers". En effet, notre but étant d'intégrer des algorithmes de classification, il est utile de savoir comment tournent les algorithmes de classifications, leur paramétrage et l'architecture pour organiser les ressources pour ces derniers.
Quelques tests ont été menés notamment pour bayes naif, nous sommes arrivé à le faire tourner via eclipse.

2.3 Ajout d'un algorithme dans Weka

//TODO

3 Prochaine étape

//TODO

1. Weka est une suite populaire de logiciels d'apprentissage automatique. Écrite en Java, développée à l'université de Waikato, Nouvelle-Zélande. Weka est un Logiciel libre disponible sous la Licence publique générale GNU.

4 Sources

//TODO