Sujet de mémoire de bachelor

Email Filtering and Data Mining

ID	IBEO1-2-10
Etudiants	José Beuret
Responsables	Dr. Olivier Biberstein
Experts	
Sujet	Le but de cette thèse de Bachelor est de développer un ensemble d'outils pour le filtrage de courriers électroniques permettant de distinguer les courriels qui concernent un domaine spécifique donné de ceux n'ayant aucun rapport avec ce dernier. La thèse propose deux approches:

- l'approche *data mining/machine learning*
- l'approche *spam filter*

La première approche consiste à déterminer les attributs pertinents qui permettront à un algorithme de *machine learning* appelé également *classifier* (réseau de Bayes, de neurones, etc.) de retenir uniquement les courriels liés à un domaine spécifique. Pour ce faire il est nécessaire de sélectionner un très large jeu de courriels préalablement classés, puis d'entraîner le *classifier* au moyen de ce jeu de données. Une fois entraîné le *classifier* est alors en mesure de classer avec une certaine précision de nouveaux courriels. La seconde approche utilise de manière similaire un *spam filter open-source* existant entraîné avec le même jeu de données; les deux classes de courriels ci-dessus correspondant aux deux classes du *spam filter*.

Objectifs principaux

- Etude de l'approche *data mining* et de l'outil *open source Weka*
- Extraction des courriels à partir de larges bases de données de courrier électroniques existantes
- Développement de l'outil de prétraitement des courriels effectuant le calcul des attributs nécessaires à l'approche *data mining*
- Entraînement et utilisation d'un *classifier* adéquat à sélectionner
- Validation qualitative et quantitative de l'approche *data-mining*

Objectifs optionnels

- Etude et sélection d'un *spam filter open source*
- Entraînement du *spam filter* avec le même jeu de données de l'approche *data mining*
- Validation et comparaison des deux approches proposées

© 2009 Haute école spécialisée bernoise Technique et informatique - Section d'informatique