

Data mining appliquée à la classification de sites web

Domaine spécialisé: Information and Communication Technologies

Chargé: Prof. Dr. Olivier Biberstein

Expert: Prof. Gilles Falquet

Notre but est d'automatiser la détection de sites Web qui vendent des produits contrefaits pour lutter contre la prolifération des contrefaçons. Pour y arriver, nous utilisons les solutions provenant du data mining. En utilisant des algorithmes d'apprentissage sur de grandes quantités de données extraites à partir de sites Web, nous pouvons générer des classifieurs capables de déterminer si un site vend des contrefaçons.

Dans ce projet, nous allons lutter contre la contrefaçon en nous focalisant sur les sites de vente de montres contrefaites. L'un des objectifs de ce projet est de développer un système capable de générer des algorithmes de classification appelés classifieurs et de les utiliser pour détecter des sites de ventes de montres contrefaites. L'autre objectif est de trouver un ou plusieurs algorithmes profitables pour classer les sites Web qui vendent des montres contrefaites.

Le prétraitement des données et les classifieurs

Un jeu de données a été constitué pour créer un échantillon des sites Web liés à la contrefaçon. Plusieurs centaines de sites ont été classés manuellement dans deux classes: les sites qui vendent des montres contrefaites et les autres sites.

Les critères sont des informations extraites des sites Web afin d'avoir une forme de données utilisables pour les algorithmes d'apprentissage. Par exemple, un critère contient le nombre d'images contenues d'un site ou la fréquence d'apparitions d'un mot. Un long travail de recherche a été effectué pour trouver les critères les plus pertinents afin d'obtenir les meilleurs classifieurs à partir des algorithmes d'apprentissage. Les critères d'un site Web sont stockés dans un vecteur, c'est-à-dire un fichier listant les différents critères.

Les fichiers ARFF (Attribute-Relation File Format)

contiennent une liste où chaque ligne représente un vecteur avec les critères qui la composent. Les vecteurs et les classes d'appartenance des sites correspondants sont réunis dans **un fichier ARFF** afin d'être utilisés par les algorithmes d'apprentissage pour générer les classifieurs.

Weka est un environnement de travail contenant une suite d'outils pour le data mining. Il comprend des algorithmes d'apprentissage et des méthodes d'évaluation de classifieurs. Nous avons utilisé les outils en ligne de commande de **Weka** pour générer et évaluer les classifieurs.

L'outil de génération de classifieurs

À partir de cela, nous avons développé un outil capable d'automatiser le processus de génération de classifieurs et d'évaluer rapidement lequel est le meilleur. L'outil est constitué de 3 applications **Python**. L'application **vector.py** extrait les critères sous la forme de vecteurs. Puis l'application **arff.py** ajoute les vecteurs déjà classés dans un fichier **training.arff** et les autres dans le fichier **instances.arff**. L'application **weka.py** génère les classifieurs dans des **fichiers .model**, les évalue et traite les vecteurs du fichier **instances.arff** avec le meilleur classifieur.

Grâce à cet outil, nous avons pu mener à bien l'étude des classifieurs en déterminant avec précision lesquels étaient les plus intéressants et profitables pour la détection de sites de vente de montres contrefaites.



Elie Jeannerat

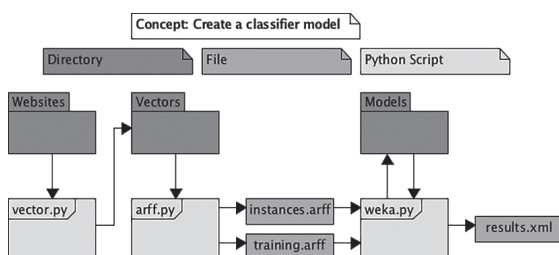


Schéma de l'outil de génération de classifieurs