Exploration de résumés personnalisés de données

Grégory Smits, Olivier Pivert

Exploration de résumés personnalisés de données

Grégory SMITS et Olivier PIVERT

Univ Rennes, CNRS, IRISA - UMR 6074
{gregory.smits | olivier.pivert}@irisa.fr

Résumé. Ce document présente une approche d'exploration et d'extraction de connaissances à partir d'un résumé linguistique et personnalisé des données.

Résumé personnalisé de données

Qu'il s'agisse d'un cadre professionnel ou personnel, le volume ainsi que l'hétérogénéité des jeux de données disponibles ne cessent d'augmenter. La valeur d'un jeu de données dépend essentiellement des connaissances que l'utilisateur peut en extraire. Un des enjeux cruciaux auxquels la communauté scientifique de gestion des données et des connaissances doit répondre concerne le développement de méthodes et d'outils permettant à un utilisateur de traduire rapidement des données brutes en connaissances interprétables et exploitables. Une manière d'accélérer ce processus de transformation des données en connaissances repose sur la génération de résumés des données, résumés qui pourront ensuite être restitués graphiquement à l'utilisateur afin qu'il dispose d'un aperçu des données. Un second axe permettant d'accélérer l'appropriation de données par un utilisateur concerne la personnalisation des représentations et des explications générées.

Une des utilisations possibles de la théorie des sous-ensembles flous consiste à transformer les domaines de définition, généralement de nature numérique ou catégorielle, des attributs qui décrivent les données en variables linguistiques. Comme l'illustre la figure 1, une variable linguistique associée à une partition floue des données permet à la fois de discrétiser un domaine de définition et également d'associer une étiquette linguistique à chaque regroupement de données. De plus, de par la possibilité de représenter des transitions graduelles entre les différents éléments de la partition, les ensembles flous constituent un cadre théorique idéal pour représenter le caractère subjectif et imprécis des termes que nous utilisons pour décrire des phénomènes observables (e.g. *'prix élevé'*, *'accélération forte'*, *'métier à risque'*, etc). Les partitions floues définies par l'utilisateur sur les attributs qui l'intéressent forment un vocabulaire utilisateur.

Nous avons conçu un algorithme distribuable dans une architecture de calcul pour quantifier dans quelle mesure chaque terme du vocabulaire utilisateur couvre un jeu de données. Le résultat de ce processus de réécriture des données forme un vecteur composé de termes linguistiques subjectifs. L'ensemble des termes linguistiques du vocabulaire qui couvrent un tant soit peu les données constitue un résumé que nous pouvons par exemple représenter graphiquement comme un nuage de mots ou une vue *ad-hoc*. La figure 1 (droite) décrit linguistiquement 127

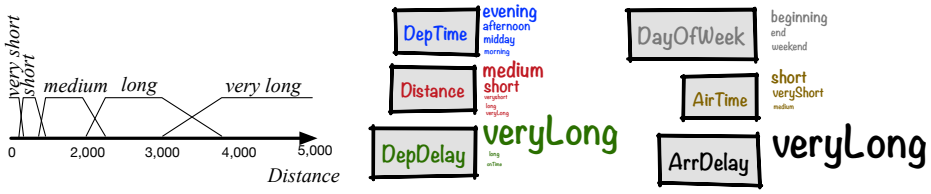

FIG. 1 – *Partition floue décrivant la durée d'un vol (gauche) et résumé des données (droite)*

millions de vols domestiques aux Etats-Unis sur six dimensions [1]. Une mesure d'information dédiée à cette représentation a également été proposée pour identifier les termes les plus informatifs d'un résumé. La représentation graphique des données est ensuite ajustée pour mettre en exergue ces termes importants.

Exploration interactive et extraction de connaissances à partir des résumés

Outre le fait de fournir une vue synthétique de grands volumes de données, la représentation graphique d'un vecteur de réécriture constitue également une interface graphique d'exploration des données. En cliquant sur un terme apparaissant dans le résumé, l'utilisateur déclenche la sélection du sous-ensemble des données satisfaisant ce terme. Ce sous-ensemble est lui-même résumé, puis représenté graphiquement afin de permettre une exploration interactive des données par combinaison conjonctive de termes linguistiques subjectifs. Nous avons construit une structure d'indexation dédiée aux vecteurs de réécriture pour permettre une navigation fluide dans les données. Nous avons également montré dans Smits et al. (2016) qu'il était possible d'extraire très efficacement des connaissances utiles (règles d'association, termes atypiques, etc.) à partir des vecteurs de réécriture construits lors du résumé des données.

Notre objectif actuel est de collaborer avec des experts en interface graphique pour construire des représentations dédiées à nos résumés linguistiques et aux connaissances que nous extrayons.

Références

Smits, G., O. Pivert, et R. R. Yager (2016). A soft computing approach to agile business intelligence. In *Fuzzy Systems (FUZZ-IEEE), 2016 IEEE International Conference on*, pp. 1850–1857. IEEE.

Summary

This short document presents a data exploration and knowledge extraction approach based on a linguistic and subjective summary of the data.

1. http ://stat-computing.org/dataexpo/2009/the-data.html