Agrégation de sac-de-sacs-de-mots pour la recherche d'information par modèles vectoriels

Vincent Claveau

IRISA – CNRS
Campus de Beaulieu, F-35042 Rennes
vincent.claveau@irisa.fr

Résumé. Cet article étudie l'intérêt de représenter les documents textuels non plus comme des sacs-de-mots, mais comme des sacs-de-sacs-de-mots. Au cœur de l'utilisation de cette représentation, le calcul de similarité entre deux objets nécessite alors d'agréger toutes les similarités entre sacs de chacun des objets. Nous évaluons cette représentation dans un cadre de recherche d'information, et étudions les propriétés attendues de ces fonctions d'agrégation. Les expériences rapportées montrent l'intérêt de cette représentation lorsque les opérateurs d'agrégation respectent certaines propriétés, avec des gains très importants par rapport aux représentations standard.

1 Introduction

La représentation sac-de-mots des documents (abrégée ici en BoW, *Bag-of-Words*) est très largement utilisée en recherche d'information (RI) et en traitement automatique des langues (TAL). Elle permet d'associer à un texte un descripteur unique basé sur l'ensemble des mots-formes qu'il contient. Cependant, cette représentation est parfois trop grossière pour certaines tâches. Plusieurs représentations alternatives ont été imaginées selon les cas et les informations disponibles. Les similarités entre objets complexes (graphes, arbres...) ont été extensivement étudiés (Bunke, 2000, *inter alia*), mais sont rarement utilisées en RI à cause de leur coût calculatoire. C'est pourquoi, dans beaucoup de cas, les travaux gardent une structure de données identique à celle des sacs-de-mots (même si ce sont des morphèmes, des n-grammes ou des syntagmes et non plus des mots qui sont manipulés). Dans cet article, nous nous intéressons à une extension simple de la représentation classique en sac-de-mots dans laquelle un objet est décrit par un multiensemble de sac-de-mots. Cette représentation en sac-de-sacs-de-mots (BoBoW, *Bag-of-Bags-of-Words* garde certaines propriétés calculatoires des BoW, mais nécessite de savoir comment agréger les résultats obtenus entre sacs-de-sacs. Dans son travail séminal en RI, Wilkinson (1994) l'utilise pour comparer une requête aux différentes portions d'un document et combiner les résultats, soit sur les similarités soit sur les rangs. Mais les quelques fonctions d'agrégation testées obtiennent des résultats inférieurs à un système vectoriel classique. En revanche, cette représentation a été utilisée avec succès dans des cadres particuliers en TAL (Ebadat et al., 2012) et en image (Kondor et Jebara, 2003). Elle est aussi à rapprocher des travaux sur la recherche d'information structurée (Luk et al., 2002) (la prise en compte du