Une nouvelle mesure pour l'évaluation des méthodes d'extraction de thématiques : la Vraisemblance Généralisée

Mohamed Dermouche*,** Julien Velcin* Sabine Loudcher* Leila Khouas**

*Laboratoire ERIC, Université Lumière Lyon2,
5 av. P. Mendès-France 69676 Bron Cedex, France
{julien.velcin, sabine.loudcher}@univ-lyon2.fr
**AMI Software R&D,
1475 av. A. Einstein 34000 Montpellier, France
{mde, lkh}@amisw.com

Résumé. Les méthodes dédiées à l'extraction automatique de thématiques sont issues de domaines variés : linguistique computationnelle, TAL, algèbre linéaire, statistique, etc. A ces méthodes spécifiques, peuvent s'ajouter des méthodes adaptées d'autres domaines, notamment de l'apprentissage automatique non supervisé. Les résultats produits par l'ensemble de ces méthodes prennent des formes hétérogènes : partitions de documents, distributions de probabilités sur les mots, matrices. Cela pose clairement un problème pour les comparer de manière uniforme. Dans cet article, nous proposons une nouvelle mesure de qualité, intitulée Vraisemblance Généralisée, pour permettre une évaluation et ainsi la comparaison de différentes méthodes d'extraction de thématiques. Les résultats, obtenus sur un corpus de documents Web autour des élections présidentielles françaises de 2012, ainsi que sur le corpus *Associated Press*, montrent la pertinence de la mesure proposée.

1 Introduction

Les documents textuels sont tellement abondants sur le Web que l'information pertinente est souvent difficile à retrouver. Dans l'objectif d'offrir une meilleure navigation dans les corpus de documents, que ce soit pour l'exploration du contenu ou la recherche d'information, l'extraction de thématiques (*topic extraction*) se distingue comme une tâche de fouille de textes dont l'objectif est d'extraire, automatiquement et sans catégories données *a priori*, des thématiques (sujets) à partir de grands corpus de documents.

L'extraction de thématiques a été étudiée par différentes communautés, que ce soit celle de la fouille de données (Anaya-Sánchez et al., 2008), du Traitement Automatique des Langues (Blei et al., 2003), de la linguistique computationnelle (Ferret, 2006) ou de la recherche d'information (Zamir et al., 1997), d'où l'existence de différentes méthodes dédiées à cette tâche. A ces méthodes spécifiques, peuvent s'ajouter des méthodes adaptées notamment de l'apprentissage automatique non supervisé. Les résultats produits par l'ensemble de ces méthodes prennent des formes hétérogènes : partitions, matrices, distributions de probabilités sur les mots, etc. Cela pose clairement un problème de comparaison de ces résultats. Dans cet article,