Regrouper les données textuelles et nommer les groupes à l'aide de classes recouvrantes

Marian-Andrei Rizoiu[*,**], Julien Velcin[*,***]
Jean-Hugues Chauchat[*,****]

[*]Laboratoire ERIC, Université Lumière Lyon2,
5 av. P. Mendès-France 69676 Bron Cedex, France
[**]Marian-Andrei.Rizoiu@univ-lyon2.fr
[***]Julien.Velcin@univ-lyon2.fr
[****]Jean-Hugues.Chauchat@univ-lyon2.fr

Résumé. Organiser les données textuelles et en tirer du sens est un défi majeur aujourd'hui. Ainsi, lorsque l'on souhaite analyser un débat en ligne ou un forum de discussion, on voudrait pouvoir rapidement voir quels sont les principaux thèmes abordés et la manière dont la discussion se structure autour d'eux. Pour cela, et parce que un même texte peut être associé à plusieurs thèmes, nous proposons une méthode originale pour regrouper les données textuelles en autorisant les chevauchements et pour nommer chaque groupe de manière lisible. La contribution principale de cet article est une méthode globale qui permet de réaliser toute la chaîne, partant des données textuelles brutes jusqu'à la caractérisation des groupes à un niveau sémantique qui dépasse le simple ensemble de mots.

1 Introduction

L'extraction d'information à partir de données non structurées, en particulier textuelles, est un domaine de recherche très actif. Internet constitue une véritable mine où l'on trouve actuellement ce type d'information : articles, blogs, *chats*, forums, débats, etc. Cette profusion explique les nombreux travaux qui cherchent à extraire le plus d'information "utile", ayant du sens, à partir de ces données. Le présent travail a été mené en étroite collaboration avec une jeune entreprise qui organise et analyse des débats en ligne.

Supposons que nous ayons un ensemble de textes qui traitent des conséquences économiques d'une décision politique. Nous aimerions disposer d'un outil capable d'extraire les principaux thèmes associés aux réactions à cette décision, et ce de manière automatique et avec un minimum de connaissance préalable sur les textes. Cet outil doit regrouper les textes puis proposer un ou plusieurs nom(s) à chacune de ces catégories. Les thématiques pourraient alors être : la "politique du gouvernement", les "propositions de l'opposition", les "réactions syndicales", l'"efficacité économique", la "justice sociale", etc. Les textes en langage naturel étant naturellement associés à plusieurs thématiques Cleuziou (2007), nous avons choisi de développer un système capable si besoin de classer un même texte dans plusieurs catégories.