Une nouvelle approche pour l'extraction non supervisée de critères

Benjamin Duthil*, François Trousset*
Mathieu Roche**, Michel Plantié*
Gérard Dray*, Jacky Montmain*

*EMA-LGI2P, Parc Scientifique Georges Besse, 30035 Nîmes Cedex, France
prénom.nom@mines-ales.fr
**LIRMM Université Montpellier 2, CNRS 5506, 161 Rue Ada, F-34392 Montpellier
prénom.nom@lirmm.fr

Résumé. Récemment de nouvelles techniques regroupées sous le vocable de détection automatique d'opinions (opinion mining) ont fait leur apparition et proposent une évaluation globale d'un document. Ainsi, elles ne permettent pas de mettre en avant le fait que les personnes expriment une opinion très positive du scénario d'un film alors qu'elles trouvent que les acteurs sont médiocres. Dans cet article, nous proposons de caractériser automatiquement les segments de textes relevant d'un critère donné sur un corpus de critiques.

1 Contexte et présentation générale

L'approche décrite dans cet article a pour objectif de réaliser une segmentation de textes selon un domaine et des critères spécifiés par l'utilisateur. Le principe général est le suivant. Dans un premier temps, l'utilisateur spécifie son domaine de recherche (e.g. cinema) et les critères sur lesquels il souhaite faire la segmentation (e.g. acteur, musique, réalisation, ...). Ensuite, pour chaque critère, il spécifie un ensemble de mots germes. Expérimentalement nous considérons que 7 mots germes sont spécifiés pour chacun des critères. Ainsi, dans le cas du critère acteur, nous considérons par exemple les mots germes (en anglais) suivants : acting, actor, casting, character, interpretation, role, star. La seconde étape consiste à utiliser un moteur de recherche (dans nos expérimentations nous avons utilisé google) pour récupérer un ensemble de textes contenant pour des documents de la thématique et pour un critère donné, au moins l'un des mots germes. A l'issue de cette étape nous disposons donc pour chacun des critères, d'un ensemble de 7 classes de documents associés à un mot germe. De la même manière nous interrogeons le moteur de recherche pour obtenir un second ensemble de documents ne contenant aucun des germes pour un critère. L'objectif de ces deux interrogations est la suivante. Dans le premier corpus (i.e. celui contenant des textes associés à un mot germe), nous souhaitons extraire des mots qui sont en corrélation avec des mots germes et ainsi étendre de manière automatique le dictionnaire de mots associés à un critère. Par exemple nous souhaitons pouvoir retrouver via une analyse de ces documents que le mot actor est souvent lié aux mots actress, interpret, player ... L'objectif du second corpus (que nous nommons anti-classe) est au contraire de pouvoir extraire l'ensemble des mots qui ne sont pas reliés au critère