

Catégorisation des évaluations dans un corpus de blogs multi-domaine

Matthieu Vernier*, Laura Monceaux*, Béatrice Daille*, Estelle Dubreil*

*LINA - CNRS UMR 6241
UFR de Sciences et Techniques, 2, rue de la Houssinière
BP 92208, 44322 Nantes Cedex 03, France
{Prénom.Nom}@univ-nantes.fr,
<http://www.lina.univ-nantes.fr>

Résumé. Dans le cadre de la fouille d'opinion, nous proposons une méthode automatique pour la détection et la catégorisation des évaluations localement exprimées dans un corpus de blogs multi-domaine. Cette méthode s'appuie sur deux théories linguistiques modélisant le processus d'évaluation dans le langage naturel et sur des ressources lexicales. Nous présentons deux composants logiciels qui s'intègrent à la plateforme *UIMA*¹ et qui permettent : l'extraction automatique de structures symboliques spécifiques à l'expression évaluative et la catégorisation des évaluations à partir des structures symboliques apprises. L'outil de catégorisation vise en particulier à analyser la signification axiologique, la modalité et la configuration énonciative d'une évaluation. L'objectif à terme est d'associer les évaluations ainsi catégorisées à leur sujet.

1 Introduction

L'image est un aspect fondamental dans notre société pour les personnes et les entreprises pour lesquelles l'opinion du public est importante. Celles-ci ont besoin de se tenir au courant de l'évolution de leur image et des sujets qui intéressent la population pour s'adapter à leurs attentes et améliorer leur réactivité. Ces aspects impliquent particulièrement l'industrie des nouvelles technologies, la politique, la publicité, les médias ou la finance pour lesquels l'étude de l'image représente un enjeu et un pouvoir économique majeur.

Les blogs se caractérisent par l'usage évaluatif qui en est fait, au sens où les internautes l'utilisent pour exprimer librement et partager leurs opinions sur leurs centres d'intérêts. De plus, la quantité de blogs et leur réactivité rendent possible des analyses en quasi temps réel et l'observation de variations au fil des jours, des semaines ou des mois. Les blogs prennent une importance de plus en plus grande ; les sociologues se penchent sur ce phénomène, d'autres médias en parlent et font référence à des informations provenant de blogs, que ce soit pour leur caractère provocateur, original, novateur, etc. De plus, ils sont en perpétuelle évolution, tant du point de vue qualitatif que quantitatif. Le nombre de blogs est passé de 8 à 72 millions

¹« Unstructured Information Management Architecture » - voir Actes de l'atelier *Towards Enhanced Interoperability for Large HLT Systems: UIMA for NLP (LREC'08)* (2008).