Extraction de commentaires utilisateurs sur le Web

Julien Subercaze*, Christophe Gravier*, Frédérique Laforest*

* Université de Lyon, F-42023, Saint-Etienne, France
CNRS, UMR5516, Laboratoire Hubert Curien, F-42000, Saint-Etienne, France,
Université de Saint-Etienne, Jean Monnet, F-42000, Saint-Etienne, France

Résumé. Dans cet article, nous présentons CommentsMiner, une solution d'extraction non supervisée pour l'extraction de commentaires utilisateurs. Notre approche se base sur une combinaison de techniques de fouille de sous-arbres fréquents, d'extraction de données et d'apprentissage de classement. Nos expérimentations montrent que CommentsMiner permet de résoudre le problème d'extraction de commentaires sur 84% d'un jeu de données représentatif et publiquement accessible, loin devant les techniques existantes d'extraction.

1 Introduction

Les grandes entreprises du Web ont depuis longtemps compris la valeur des contenus générés par leur utilisateurs – des services gratuits de grande qualité leur permettent d'atteindre plusieurs millions d'utilisateurs. Analyser et traiter ces données a fait émerger plusieurs défis scientifiques, il n'est donc pas surprenant de constater que ces contenus attirent les recherches en informatique. D'autres données sociales du même acabit, cette fois-ci disséminées sur la toile, sont les commentaires laissés par les utilisateurs du Web. Dans cet article, nous proposons CommentsMiner, un algorithme en deux temps permettant l'extraction des commentaires présents dans une page, ainsi que leur structure conversationnelle. Notre approche permet l'extraction de commentaires en réponse à des commentaires déjà publiés : l'extraction de la structure conversationnelle est fondamentale pour les tâches d'analyse en aval. Le problème d'extraction de commentaires utilisateurs peut être défini comme retrouver un motif p ayant été utilisé afin de générer le fragment HTML utilisé pour générer les commentaires dans la page. Ce motif est valide seulement à l'échelle d'un site Web : une fois le motif déterminé pour la page d'un site, la probabilité d'extraire correctement les commentaires présents dans les autres pages du site est grande.

La figure 1 présente l'architecture générale de CommentsMiner. Dans un premier temps, le problème d'extraction des commentaires est vu comme une tâche de fouille sous contraintes des sous-arbres fréquents induits et clos. Cette étape renvoie des motifs $p_i \in P$ récurrents dans la page (motif pour générer des commentaires, mais aussi les menus, publicités, contenus divers, etc.) et donc candidats pour être ceux utilisés pour générer le fragment HTML représentant les commentaires. Dans un second temps, l'apprentissage d'une fonction de classement permet de sélectionner le motif p étant le plus probablement celui utilisé pour présenter des commentaires dans la page. Nous présentons successivement ces étapes avant de présenter nos résultats expérimentaux dans la dernière section de cet article.