Proxem Ubiq : une solution d'e-réputation par analyse de feedbacks clients

François-Régis Chaumartin
Proxem, 19 bd de Magenta, 75010 Paris
frc@proxem.com

Mots-clés : e-réputation, reconnaissance d'entités nommées, classification, clustering, analyse syntaxique, apprentissage

Être à l'écoute de ses clients est un enjeu majeur pour toute grande marque. Les verbatims d'expression spontanée des consommateurs se trouvent le plus souvent sur des sources externes (blogs, forums, news, RSS, tweets…) et internes (mails envoyés spontanément, réponses aux questions ouvertes de sondages). Ubiq permet aux entreprises de calculer leur e-réputation en analysant ces différents feedbacks. Ubiq identifie les attentes des consommateurs, détecte les tendances, analyse les opinions et permet d'anticiper des problèmes. En un coup d'œil, on visualise les « sujets chauds » du moment.

La plateforme de TAL Antelope est au cœur d'Ubiq. L'analyse sémantique effectuée enchaîne plusieurs opérations.

(1) La qualité des documents traités étant très variable, une correction orthographique est souvent nécessaire ; néanmoins, cette opération doit être effectuée avec une connaissance du contexte métier ; par exemple, les noms de marques qui viennent d'apparaître (et ne figurent pas encore dans un lexique) ne doivent pas être « corrigés » vers un mot proche.

(2) La reconnaissance d'entités nommées vise classiquement à identifier des personnes, lieux et organisation. Dans un contexte d'enseigne de grande distribution, les entités intéressantes à détecter sont plutôt les produits, marques et concurrents cités, ainsi que des concepts liés au métier (le risque sanitaire ou le risque juridique, par exemple). Nous avons développé une nouvelle approche d'acquisition à large échelle d'entités nommées. (2a) Une première phase d'extraction terminologique permet d'amorcer la liste des concepts du domaine. (2b) Une seconde phase utilise deux ressources de large couverture (la Wikipédia et un WordNet pour le français) pour créer des gazettes ; en cas d'ambiguïté possible (*orange* fruit ou *Orange* marque), les termes des gazettes sont automatiquement associés à des mots clés activateurs ou inhibiteurs (pour les deux sens d'orange : jus, fruit, pulpe… ou internet, contrat, carte sim, opérateur…). (2c) L'application de ces gazettes permet de constituer un premier corpus annoté selon les entités nommées du domaine. Un apprentissage (par CRF) est alors effectué sur le corpus, pour identifier de nouvelles instances d'entités. (2d) Chaque document fait aussi l'objet d'une classification multi-motifs (dont une analyse d'opinion pour en déterminer la valence).

(3) L'ensemble des documents est partitionné en sous-ensemble homogènes, pour déterminer les tendances du moment ; l'utilisation de techniques de clustering spectral permet de traiter en quelques minutes plusieurs milliers de documents.

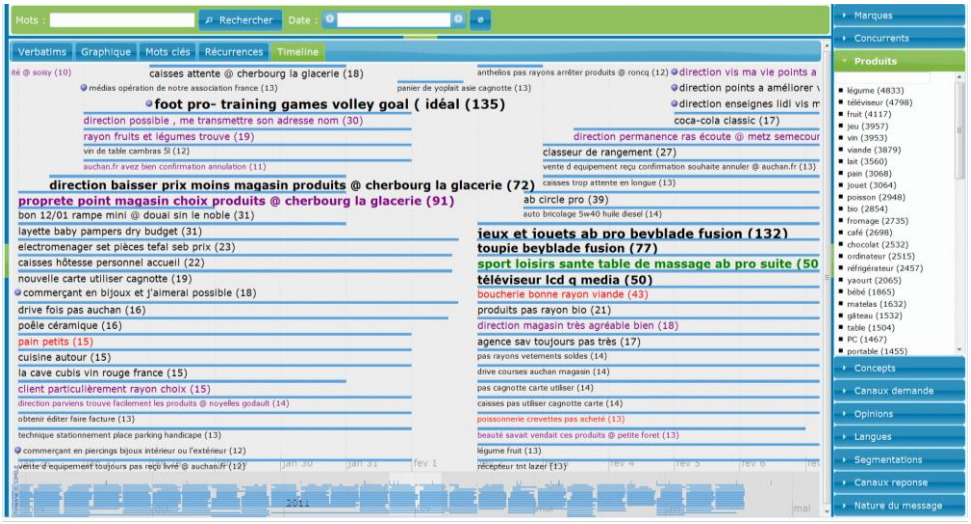

Figure 1 : Une capture d'écran d'Ubiq, montrant d'une façon synthétique ce qui s'est passé pendant deux semaines dans une enseigne de la grande distribution. La partie centrale est le « résumé sémantique » de plus de 10 000 feedbacks.