Discours de presse et veille stratégique d'événements

Approche textométrique et extraction d'informations pour la fouille de textes

Résumé

Ce travail a pour objet l'étude de deux méthodes de fouille automatique de textes, l'extraction d'informations et la textométrie, toutes deux mises au service de la veille stratégique des événements économiques. Pour l'extraction d'informations, il s'agit d'identifier et d'étiqueter des unités de connaissances, entités nommées — *sociétés, lieux, personnes*, qui servent de points d'entrée pour les analyses d'activités ou d'événements économiques — *fusions, faillites, partenariats*, impliquant ces différents acteurs. La méthode textométrique, en revanche, met en œuvre un ensemble de modèles statistiques permettant l'analyse des distributions de mots dans de vastes corpus, afin faire émerger les caractéristiques significatives des données textuelles. Dans cette recherche, la textométrie, traditionnellement considérée comme étant incompatible avec la fouille par l'extraction, est substituée à cette dernière pour obtenir des informations sur des événements économiques dans le discours. Plusieurs analyses textométriques (spécificités et cooccurrences) sont donc menées sur un corpus de flux de presse numérisé. On étudie ensuite les résultats obtenus grâce à la textométrie en vue de les comparer aux *connaissances* mises en évidence au moyen d'une procédure d'extraction d'informations. On constate que chacune des approches contribuent différemment au traitement des données textuelles, produisant toutes deux des analyses complémentaires. À l'issue de la comparaison est exposé l'apport des deux méthodes de fouille pour la veille d'événements.

Mots-clés : textométrie, extraction d'informations, événements, veille stratégique, fouille de textes, discours de presse, spécifictés, cooccurrences

News Discourse and Strategic Monitoring of Events

Textometry and Information Extraction for Text Mining

Abstract

This research demonstrates two methods of text mining for strategic monitoring purposes: information extraction and Textometry. In strategic monitoring, text mining is used to automatically obtain information on the activities of corporations. For this objective, information extraction identifies and labels units of information, named entities (*companies, places, people*), which then constitute entry points for the analysis of economic activities or events. These include mergers, bankruptcies, partnerships, etc., involving corresponding corporations. A Textometric method, however, uses several statistical models to study the distribution of words in large corpora, with the goal of shedding light on significant characteristics of the textual data. In this research, Textometry, an approach traditionally considered incompatible with information extraction methods, is applied to the same corpus as an information extraction procedure in order to obtain information on economic events. Several textometric analyses (characteristic elements, co-occurrences) are examined on a corpus of online news feeds. The results are then compared to those produced by the information extraction procedure. Both approaches contribute differently to processing textual data, producing complementary analyses of the corpus. Following the comparison, this research presents the advantages for these two text mining methods in strategic monitoring of current events.

Keywords: textometry, information extraction, events, business intelligence, text mining, news discourse, characteristic elements, co-occurrences

UNIVERSITE SORBONNE NOUVELLE - PARIS 3

268 Langage et langues

SYLED CLA^{2}T

19 rue des Bernardins 75019 PARIS