
Introduction générale sur les enjeux du Text et Data Mining

Patrice Bellot*¹

¹Chargé de mission IST/Fouille de données textuelles à l'institut des sciences de l'information et de leur interactions – (INS2I) – France

Résumé

Patrice Bellot est professeur en informatique à Aix-Marseille Université, membre du Laboratoire d'Informatique et Systèmes (LIS - CNRS/Aix-Marseille Université/Université de Toulon) et de l'école d'ingénieurs Polytech. Spécialiste de la recherche d'information et du traitement automatique des langues, il a soutenu une thèse de doctorat en 2000 sur les méthodes de classification et de segmentation thématiques de textes. Maître de conférences à l'Université d'Avignon entre 2000 et 2011, il a depuis rejoint Aix-Marseille Université.

Patrice Bellot est responsable depuis sa création en 2013 de l'équipe DIMAG " Data, Information & content Management Group " du LIS. Il est par ailleurs directeur scientifique de l'OpenEdition Lab, structure de recherche au sein de l'équipement d'excellence (Equipex) et Infrastructure Nationale OpenEdition portée par le CNRS, Aix-Marseille Université, l'EHESS et l'Université d'Avignon. Lauréat en 2011 et 2012 d'un Google Research Award en humanités numériques dans le domaine de l'extraction d'information au sein d'articles scientifiques, ses recherches actuelles portent sur la fouille de textes, l'analyse de sentiments et les méthodes numériques de recherche d'information et de recommandation automatique de contenus.

Résumé de l'intervention

La fouille de données et de textes (Text & Data Mining, TDM) est un domaine apparu il y a une trentaine d'années et qui est désormais associée à de nombreux enjeux scientifiques, industriels et sociétaux. À l'intersection de l'analyse de données, de l'intelligence artificielle et du Traitement Automatique des Langues, le TDM est impliqué dans de très nombreuses applications utilisées au quotidien par les chercheurs, les entreprises et le grand public. Citons par exemple la recherche d'information et la recommandation automatique de contenus, la navigation au sein de bibliothèques numériques et le filtrage d'informations. Ces applications ont en commun quantité de traitements génériques sur les données, qu'elles soient brutes ou textuelles, structurées ou non.

Cette introduction aux enjeux du TDM sera illustrée par quelques exemples soulignant les actuels verrous technologiques et scientifiques mais aussi la nécessité d'œuvrer de façon commune au développement de logiciels et de services ouverts. L'objectif est de permettre le développement et l'appropriation de plateformes aptes à fouiller l'immense production scientifique, à la lumière des services et APIs ISTEEX, mais aussi de rendre transparents les

*Intervenant

traitements qui sont au cœur d'activités quotidiennes.

Vidéo de l'intervention (lien à copier/coller) : <https://webcast.in2p3.fr/video/introduction-generale-sur-les-enjeux>