

De la statistique des données à la statistique des connaissances : avancées récentes en analyse des données symboliques

Edwin Diday

Université Paris-Dauphine
CEREMADE,
Place Du Maréchal de Lattre de Tassigny.
75016 Paris, FRANCE
adresse@email
<http://www.une-page.html>

Il s'agit de montrer que l'Extraction et la Gestion des Connaissances (EGC) n'est pas seulement un domaine d'utilisation normale des outils standards de la Statistique, du Data Mining ou de l'Analyse des Données plus ou moins complexes, mais que de plus, le fait de s'intéresser aux connaissances et aux concepts qui en forment les atomes en tant qu'unités d'étude remet totalement en cause ces outils et nécessite leur renouvellement complet aussi bien dans leur théorie que dans leur pratique et dans la façon de les penser.

Partout dans le monde se réalisent maintenant des bases de données de taille parfois gigantesque. Ces bases constituent des gisements de connaissances qu'il est possible voire indispensable d'extraire aussi bien dans le domaine de la gestion industrielle que dans celui de l'administration ou de la recherche. De ces bases on peut extraire des variables qualitatives ou des produits cartésiens de telles variables qui définissent des catégories réifiées en concepts constituant ainsi un nouveau type d'unité statistique que l'on peut décrire, modéliser et analyser en terme d'intension et d'extension. Pour tenir compte de la variation interne aux concepts, on est conduit à utiliser des variables non purement numériques (donc dites " symboliques ") à valeur intervalle, histogramme, fonction ou ensemble muni parfois de règles et de taxonomies. L'Analyse des Données Symboliques et son logiciel SODAS* (issu de deux projets Européens soutenus par EUROSTAT où 18 équipes de 9 pays ont œuvrés ensemble) a pour but d'étendre la Statistique, le Data Mining et l'Analyse des Données classiques ou complexes aux concepts pour extraire de nouveaux concepts qui peuvent être à leur tour analysés.

Références

- L. Billard, E. Diday (2003) "From the statistic of Data to the statistic of Knowledge". JASA. Journal of the American Statistical Association. Juin.
Bock H.H., Diday E. (2000) "Analysis of Symbolic Data". Study in Classification, Data Analysis and Knowledge Organisation. Springer Verlag.

Electronical Journal of Symbolic Data Analysis (JSDA)
www.jsda.unina2.it/newjsda/volumes/index.htm

- Informations, formation et chargement de SODAS

* www.ceremade.dauphine.fr/%7Etuati/sodas-pagegarde.htm.