Essai de Typologie Structurelle des Indices de Similarité Vectoriels par Unification Relationnelle

François Marcotorchino

Thales Communications : 160, boulevard de Valmy – BP 82
92704 Colombes Cedex et Laboratoire de Statistique Théorique et Appliquée (Paris VI)
jeanfrancois.marcotorchino@fr.thalesgroup.com

Résumé : Cet article a pour but de proposer un regard nouveau et unificateur à la problématique des Indices de Similarité et des Critères de structuration ou de partitionnement. Une catégorisation des indices, des propriétés non connues ainsi qu'une présentation dans différents axes de structuration seront suggérées. La recherche des significations et des filiations associées sera donnée comme résultat dérivé de ce travail.

1 Introduction

La recherche sur les indices de similarité a, depuis longtemps, donné lieu à une abondante littérature, les références aux indices ayant été faites souvent, même dans les meilleurs articles, sous forme de listes (type inventaire) sans qu'une réelle structuration n'ait été proposée. Nombre de ces indices ont été introduits et donnés dans différents articles et dans différents domaines fondamentaux ou d'application, au fur et à mesure de leur utilisation potentielle. Ainsi n'est-il pas étonnant de trouver ces indices proposés et introduits dans des domaines aussi variés que: *les Sciences Humaines* et plus particulièrement la Sociologie et l'Ethnologie et ses dérivées : Ethnopsychologie, Ethnogénétique, etc.., *la Linguistique* proprement dite, dont: Lexicologie, Lexicométrie, Ethnolinguistique, Text Mining, *les Mathématiques*: Analyse des données, Classification et « Clustering », *les Sciences du vivants* : la Biologie, la Biométrie, la Physiologie, la Phylogénie, la Zoologie, la Médecine, *les Sciences organiques* : la Chimie moléculaire, la Biochimie et enfin de nombreux domaines plus « business », comme : le « Customer Relationship Management », le « Business Intelligence », la « Géo-cartographie », le «Profiling » etc..

Presque tous les indices courants ont été introduits à des périodes et à des dates différentes, pour des buts et motifs variés sans structuration et explication claire du rôle de chacun et sans aucun regard sur une filiation ou une hérédité sous jacentes permettant de mieux les comprendre ou de les interpréter (ceci étant sans doute dû à la provenance très différenciée des inventeurs). Ce phénomène s'est traduit, de fait, soit par une impression de fouillis, soit, et on le verra plus loin dans ce document, par des successions de redécouvertes (parfois très récentes) d'indices existants depuis fort longtemps, ou de mises en évidence de propriétés connues depuis très longtemps par des chercheurs de disciplines différentes.

Preuve que le sujet est toujours d'actualité, un article vraiment très récent de Matthijs J. Warrens (2009) vient de paraître dans la Revue Journal of Classification, au moment