Journées de méthodologie statistique de l'Insee 2022

Quelques limites de l'algorithme implémenté dans l'outil Sicore

Théo LEROY (*)

(*) Insee, Direction de la méthodologie et de la coordination statistique et internationale

theo.leroy@insee.fr

Mots-clés. (6 maximum) : traitement du langage naturel, arbres de décision, moteur de règles, entropie

Domaines. Codification automatique

Résumé

Sicore (Système Informatique de Codage de Réponse aux Enquêtes) est, depuis 1993, le principal logiciel de codification automatique de l'Insee. Il permet de coder des libellés dans des nomenclatures. Son utilisation s'est imposée sur une large variété de données : professions,activités d'entreprise, diplômes, communes, nationalités, produits... Sicore intervient dans de nombreuses productions de l'institut (le recensement de la population, les enquêtes ménages, le répertoire Sirene...). En dépit, de ces applications, les choix méthodologiques adhérents à cet outil ne sont pas toujours évidents.

La codification avec Sicore se déroule en 3 parties. Une première étape consiste à normaliser le libellé à coder pour simplifier sa manipulation par la suite. Des caractères sont alors supprimés et des règles de synonymisation sont appliquées. Une seconde phase vise à retrouver le libellé normalisé au sein d'un index de libellé de référence selon un arbre de décision questionnant des groupes de caractères consécutifs appelés « atomes » qui composent le libellé. Enfin, si le libellé est retrouvé dans l'index et si le libellé à lui seul n'est pas suffisant pour coder dans la nomenclature, des variables annexes catégorielles sont prises en compte au travers de règles logiques. Sicore est un système expert où tous les composants de l'algorithme (règles de normalisation, index de références, règles logiques) doivent être spécifiés en intégralité par un ou plusieurs experts de la nomenclature. Cet article discute de deux limites autour de cette méthode.

La première partie de l'article interroge la pertinence des arbres de décision Sicore pour rechercher un libellé dans un index. Ce type de classifieur malgré sa grande simplicité ne va pas nécessairement de soi pour répondre à un besoin codification dans une nomenclature car il est souvent associé à des risques de sur-apprentissage. En effet, les arbres de décision Sicore sont construits à partir de l'index sur des critères d'entropie locaux. Ainsi, chaque noeud de l'arbre est créé de telle sorte que le désordre dans les codes de la nomenclature possibles depuis ce

noeud soit minimal. Ce choix d'optimum conduit à des arbres assez peu profonds où, parfois, seuls quelques atomes interviennent dans la codification. Afin d'augmenter la fiabilité (pour que la décision d'attribuer un code ne soit pas fondée sur un nombre trop réduit de caractères), l'expert de la codification peut spécifier des atomes additionnels (dits de « redondance ») où la conformité devra être vérifiée. Ce phénomène est très fréquemment employé en pratique mais tend à rendre l'utilisation d'arbre pour inférer une méthode d'identification du libellé dans un index inutile en ajoutant des contrôles. Il s'agira d'évaluer pour différents types de libellés (communes, professions, activités...) à quel point l'arbre de décision avec prise en compte des atomes de redondance est plus efficace c'est à dire qu'il permet de coder plus de libellé qu'une méthode naïve d'appariement strict avec l'index.

Dans une deuxième partie, il s'agira de mettre en évidence des limites de l'approche système expert pour la codification automatique de libellés issus d'enquêtes ou de sources administratives. Le bon fonctionnement de Sicore repose sur la capacité à extraire les connaissances d'experts et à les formaliser sous forme de règles. Ce type de solution peut entraîner des problèmes de maintenance ou de gestion du volume. Au fil du temps, malgré les compétences et la volonté des experts, il devient difficile d'ajouter des nouvelles règles sans engendrer des effets de bords car le moteur de règles devient trop complexe. Par ailleurs, il est impossible pour l'expert d'intégrer toute la reprise manuelle sous forme de nouvelles règles. Le but sera de donner des ordres de grandeur de l'évolution du nombre de règles et du nombre de libellé dans l'index de référence d'une version à l'autre d'une base de connaissances et d'essayer d'évaluer la répercussion de ces mises à jour sur l'efficacité de la codification automatique.

Bibliographie

[1] RIVIERE Pascal, Sicore, un outil et une méthode pour le chiffrement automatique à l'Insee", Courrier des statistiques, n°74 - août 1995

[2] LORIGNY Jacques, QUID, une méthode générale de chiffrement automatique, Techniques d'enquêtes, Vol 14, Décembre 1988