

FC-SWEEPER : Un outil d'extraction et de navigation dans les top-k concepts formels

Amira Mouakher*

*Laboratoire Connaissance et Intelligence Artificielle Distribuées (CIAD)
Université Bourgogne Franche-Comté, 21000 Dijon, France
amira.mouakher@u-bourgogne.fr

1 Présentation de l'approche

Avec l'émergence du numérique de la dernière décennie, nous faisons face à de grands volumes de données issus de plusieurs secteurs d'activité tels que Commerce, Biologie, Médecine, Télécommunication, etc. Ces différents jeux de données véhiculent une quantité d'informations prodigieuses et pertinentes. Cependant, l'exploitation optimale de ces masses de données reste encore difficile. Ainsi, la mise en place de nouvelles solutions d'analyse de données est devenue un véritable défi pour la communauté scientifique.

Dans ce contexte, la batterie de résultats fournie par l'Analyse Formelle des Concepts s'avère d'une grande importance dans le processus d'extraction de connaissances à travers les treillis de Galois. Cependant, leur vraie exploitation a été toujours freinée par le nombre exorbitants des concepts formels extraits. Dans le but de filtrer ces derniers, plusieurs approches ont été définies. Parmi ces approches, nous nous intéressons à celles qui ont utilisé les métriques de qualité pour garder les concepts les plus intéressants.

Plusieurs mesures de qualité ont été proposées dans la littérature telles que la stabilité définie par (Kuznetsov (1990))¹, le couplage et la cohésion proposés par (Paul et Scott (2008)), la séparation introduite par (Klimushkin et al. (2010)), la distance définie par (Eklund et al. (2012)), etc. Ce nombre important engendre de nouveaux problèmes, entre autres, le choix de mesures de qualité à utiliser.

Cet article propose une nouvelle approche multi-critères permettant de sélectionner les top-k meilleurs concepts d'un ensemble de concepts formels, ayant en entrée un ensemble de mesures de qualité. En outre, il importe de souligner que leur importance peut être éventuellement pondérée par un utilisateur et/ou expert.

Parmi les méthodes multi-critères, nous proposons d'utiliser l'approche TOPSIS (Technique for Order of Preference by Similarity to Ideal Solution) introduite par (Hwang et Yoon (1981)) pour générer les top-k concepts formels. L'idée fondamentale de cette méthode consiste à choisir une solution, qui se rapproche le plus de la solution idéale (meilleure sur tous les critères) et de s'éloigner le plus possible de la pire solution (qui dégrade tous les critères).

1. Dans cette approche, nous appliquons l'algorithme DFSP proposé par (Mouakher et Ben Yahia (2019)) pour le calcul de la stabilité. Ce dernier ne nécessite aucune relation d'ordre entre les concepts.

Dans le but d'illustrer cette approche, un prototype appelé FC-SWEEPER a été développé. Ce dernier permet à un utilisateur de sélectionner un jeu de données décrivant un contexte formel et de construire le treillis de concepts. Cette construction peut se faire à travers l'énumération de tous les concepts formels qui peuvent être générés du contexte et ceci en appliquant l'algorithme LCM (Uno et al. (2004)). Dans le cas où le treillis s'avère très dense, l'utilisateur peut choisir d'extraire une couverture de concepts formels en se basant sur l'algorithme QUALITYCOVER (Mouakher et Ben Yahia (2016)) ou GRECOND (Belohlavek et Vychodil (2009)). Par la suite, l'utilisateur définit les critères à maximiser ou à minimiser, lance l'exécution de la méthode TOPSIS et visualise le résultat des top-k concepts formels.

Nous envisageons de poursuivre cette approche en l'améliorant selon deux directions : la proposition de nouveaux algorithmes de génération des top k concepts, capables de passer à l'échelle sur des jeux de données de taille réelle ainsi que d'autres méthodes d'agrégation plus fines qui ne nécessitent pas de fixer au préalable le poids des critères, e.g. les algorithmes d'extraction des motifs non dominés (Bouker et al. (2014)).

Références

- Belohlavek, R. et V. Vychodil (2009). Discovery of optimal factors in binary data via a novel method of matrix decomposition. *JCSS* 76(1), 3–20.
- Bouker, S., R. Saidi, S. Ben Yahia, et E. Mephu Nguifo (2014). Mining undominated association rules through interestingness measures. *Int. J. on Artif. Intell. Tools* 23(4).
- Eklund, P. W., J. Ducrou, et F. Dau (2012). Concept similarity and related categories in information retrieval using formal concept analysis. *IJGS* 41(8), 826–846.
- Hwang, C. L. et K. Yoon (1981). *Multiple attribut decision making : Methods and applications*. Springer-Verlag.
- Klimushkin, M., S. A. Obiedkov, et C. Roth (2010). Approaches to the selection of relevant concepts in the case of noisy data. In *Proceedings of the 8th International Conference (ICFCA), Agadir, Morocco*, Volume 5986, pp. 255–266. Springer.
- Kuznetsov, S. O. (1990). Stability as an estimate of the degree of substantiation of hypotheses derived on the basis of operational similarity. *Automatic Documentation and Mathematical Linguistics* 24(6), 62–75.
- Mouakher, A. et S. Ben Yahia (2016). Qualitycover : Efficient binary relation coverage guided by induced knowledge quality. *Inf. Sci.* 355-356, 58–73.
- Mouakher, A. et S. Ben Yahia (2019). On the efficient stability computation for the selection of interesting formal concepts. *Inf. Sci.* 472, 15–34.
- Paul, G. et D. Scott (2008). New coupling and cohesion metrics for evaluation of software component reusability. In *Proceedings of The 9th International Conference for Young Computer Scientists*, Hunan, China, pp. 1181–1186.
- Uno, T., T. Asai, Y. Uchida, et H. Arimura. (2004). An efficient algorithm for enumerating closed patterns in transaction databases. In *Proceedings of the 7th International conference Discovery Science*, Padova, Italy, pp. 16–31.