Classification automatique de mesures de similarité

Fabien Duchateau, Franck Favetta (prénom.nom@liris.cnrs.fr)

Contexte. L'appariement d'entités consiste à identifier des objets équivalents dans des jeux de données [1]. Par exemple, dans les tableaux ci-dessous, l'objet *123* du premier jeu de données correspond à l'objet *1* dans le second. Pour détecter ces objets correspondants, on utilise des mesures de similarité qui calculent un score entre les propriétés des objets à comparer. Par exemple, le coefficient de Dice calcule un score de 0,51 entre les chaînes *Holy Grail* et *Holy Grail - Dark Passenger*. Il existe de nombreuses mesures de similarité [2, 3, 4], qui sont généralement catégorisées selon leur fonctionnement. Dans ce projet, nous voulons classer les mesures selon leurs résultats.

id	nom	type
123	Les deux tours	livre
345	Holy grail	DVD
567	Les deux tours	DVD

id	titre	catégorie
1	Le seigneur des anneaux - Les 2 tours	roman
2	Holy Grail - Dark Passenger	musique

Objectifs. Ce projet cherche à atteindre les objectifs suivants :

- Développer un script qui calcule des scores de similarité, pour un ensemble de mesures et un jeu de données, puis appliquer des algorithmes de classification (clustering) pour catégoriser ces mesures.
- Étudier les résultats obtenus (comparaison avec la classification traditionnelle, robustesse de la classification, configuration des paramètres comme le nombre de catégories, etc.).

Contraintes et compétences. Lecture d'articles en anglais, programmation en Python.

Bibliographie

[1] Wei Shen, Jianyong Wang, and Jiawei Han. Entity linking with a knowledge base: Issues, techniques, and solutions. *Knowledge and Data Engineering, IEEE Transactions on*, 27(2):443–460, 2015. Lien TKDE14.

[2] Seung seok Choi and Sung hyuk Cha. A survey of binary similarity and distance measures. *Journal of Systemics, Cybernetics and Informatics*, pages 43–48, 2010. Lien JSCI2010.

[3] Nick Koudas, Sunita Sarawagi, and Divesh Srivastava. Record linkage: similarity measures and algorithms. In *SIGMOD*, pages 802–803. ACM, 2006. Lien SM2006.

[4] Najlah Gali, Radu Mariescu-Istodor, and Pasi Fränti. Similarity measures for title matching. In *2016 23rd International Conference on Pattern Recognition*, pages 1548–1553. IEEE, 2016. Lien ICPR2016.