Tang, Jiliang; **Wang, Xufei**; **Gao, Huiji**; **Hu, Xia**; **Liu, Huan**
Enriching short text representation in microblog for clustering. (English) Zbl 1251.68303
Front. Comput. Sci. 6, No. 1, 88-101 (2012).

Summary: Social media websites allow users to exchange short texts such as tweets via microblogs and user status in friendship networks. Their limited length, pervasive abbreviations, and coined acronyms and words exacerbate the problems of synonymy and polysemy, and bring about new challenges to data mining applications such as text clustering and classification. To address these issues, we dissect some potential causes and devise an efficient approach that enriches data representation by employing machine translation to increase the number of features from different languages. Then we propose a novel framework which performs multi-language knowledge integration and feature reduction simultaneously through matrix factorization techniques. The proposed approach is evaluated extensively in terms of effectiveness on two social media datasets from Facebook and Twitter. With its significant performance improvement, we further investigate potential factors that contribute to the improved performance.

MSC:

68U15 Computing methodologies for text processing; mathematical typography
68T30 Knowledge representation
91D30 Social networks; opinion dynamics

Keywords:

short texts; text representation; multi-language knowledge; matrix factorization; social media

Full Text: DOI