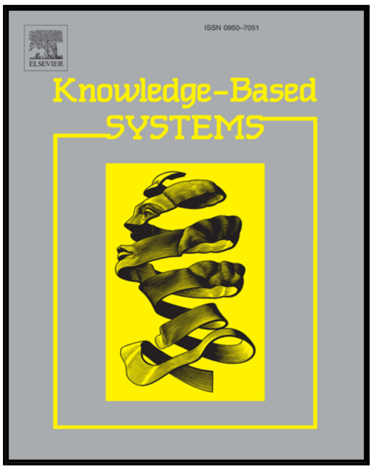

A novel regularized concept factorization for document clustering

Wei Yan, Bob Zhang, Sihan Ma, Zuyuan Yang

PII: S0950-7051(17)30367-2
DOI: 10.1016/j.knosys.2017.08.010
Reference: KNOSYS 4008

To appear in: *Knowledge-Based Systems*

Received date: 27 March 2017
Revised date: 25 July 2017
Accepted date: 9 August 2017

A novel regularized concept factorization for document clustering

Wei Yan[a], Bob Zhang[a,*], Sihan Ma[b], Zuyuan Yang[c]

[a] *Department of Computer and Information Science, University of Macau, Macau, China*
[b] *Department of Computer and Information Science, Wuhan University, Wuhan, China*
[c] *School of Automation, Guangdong University of Technology, Guangzhou, China*

Abstract

Document clustering is an important tool for text mining with its goal in grouping similar documents into a single cluster. As typical clustering methods, Concept Factorization (CF) and its variants have gained attention in recent studies. To improve the clustering performance, most of the CF methods use additional supervisory information to guide the clustering process. When the amount of supervisory information is scarce, the improved performance of CF methods will be limited. To overcome this limitation, this paper proposes a novel regularized concept factorization (RCF) algorithm with dual connected constraints, which focuses on whether two documents belong to the same class (must-connected constraint) or different classes (cannot-connected constraint). RCF propagates the limited constraint information from constrained samples to unconstrained samples, allowing the collection of constraint information from the entire data set. This information is used to construct a new data similarity matrix that concentrates on the local discriminative structure of data. The similarity matrix is incorporated as a regularization term in the CF objective function. By doing so, RCF is able to make full use of the supervisory information to preserve the local structure of the data set. Thus, the clustering performance will be improved significantly. Our experiments on standard document databases

[*] Corresponding author
Email addresses: helloyanwei@163.com (Wei Yan), bobzhang@umac.mo (Bob Zhang), sihanma@whu.edu.cn (Sihan Ma), yangzuyuan@gdut.edu.cn (Zuyuan Yang)