

## Accepted Manuscript

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](https://doi.org/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

Please cite this article as: Wei Yan, Bob Zhang, Sihan Ma, Zuyuan Yang, A novel regularized concept factorization for document clustering, *Knowledge-Based Systems* (2017), doi: [10.1016/j.knosys.2017.08.010](https://doi.org/10.1016/j.knosys.2017.08.010)

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

# A novel regularized concept factorization for document clustering

Wei Yan<sup>a</sup>, Bob Zhang<sup>a,\*</sup>, Sihan Ma<sup>b</sup>, Zuyuan Yang<sup>c</sup>

<sup>a</sup>*Department of Computer and Information Science, University of Macau, Macau, China*

<sup>b</sup>*Department of Computer and Information Science, Wuhan University, Wuhan, China*

<sup>c</sup>*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](mailto:helloyanwei@163.com) (Wei Yan), [bobzhang@umac.mo](mailto:bobzhang@umac.mo) (Bob Zhang), [sihanma@whu.edu.cn](mailto:sihanma@whu.edu.cn) (Sihan Ma), [yangzuyuan@gdut.edu.cn](mailto:yangzuyuan@gdut.edu.cn) (Zuyuan Yang)

Download English Version:

<https://daneshyari.com/en/article/4946058>

Download Persian Version:

<https://daneshyari.com/article/4946058>

[Daneshyari.com](https://daneshyari.com)