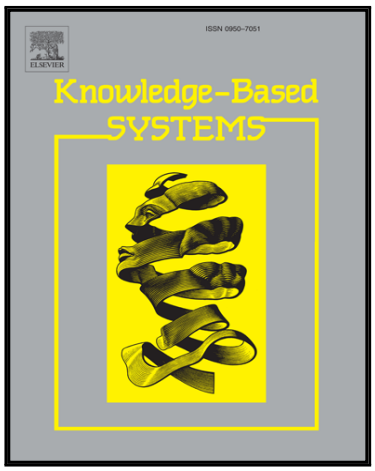

Using Bipartite Heterogeneous Networks to Speed Up Inductive Semi-Supervised Learning and Improve Automatic Text Categorization

Rafael Geraldeli Rossi, Alneu de Andrade Lopes,
Solange Oliveira Rezende

PII: S0950-7051(17)30290-3
DOI: 10.1016/j.knosys.2017.06.016
Reference: KNOSYS 3944

To appear in: *Knowledge-Based Systems*

Received date: 15 September 2016
Revised date: 5 June 2017
Accepted date: 8 June 2017

Please cite this article as: Rafael Geraldeli Rossi, Alneu de Andrade Lopes, Solange Oliveira Rezende, Using Bipartite Heterogeneous Networks to Speed Up Inductive Semi-Supervised Learning and Improve Automatic Text Categorization, *Knowledge-Based Systems* (2017), doi: 10.1016/j.knosys.2017.06.016

ELSEVIER

Available online at www.sciencedirect.com

ScienceDirect

Knowledge-Based Systems 00 (2017) 1–24

KNOSYS

Using Bipartite Heterogeneous Networks to Speed Up Inductive Semi-Supervised Learning and Improve Automatic Text Categorization

Rafael Geraldeli Rossi[a], Alneu de Andrade Lopes[b], Solange Oliveira Rezende[b]

[a] *Federal University of Mato Grosso do Sul - Três Lagoas Campus*
Ranulpho Marques Leal, 3484, ZIP 79620-080, P.O. Box 210, Três Lagoas, MS, Brazil
[b] *Institute of Mathematics and Computer Science - University of São Paulo*
Avenida Trabalhador São Carlense, 400, ZIP 13566-590, P.O. Box 668, São Carlos, SP, Brazil

Abstract

Due to the volume of texts available in digital form, the organization, management and knowledge extraction are laborious and frequently impossible to be handled. To automatically cope with these tasks, usually classification models are generated through supervised learning techniques. Unfortunately, this type of learning usually demands a huge human effort to label large volume of texts to build accurate classification models. Since collecting unlabeled texts is easy and inexpensive in several domains, the generation of classification models through inductive semi-supervised learning has been highlighted in recent years. Inductive semi-supervised learning allows to build a classification model using labeled and unlabeled texts. In this scenario, the goal is to augment the set of labeled documents with unlabeled documents to better discriminate class patterns. Hence, fewer texts must be previously labeled. However, semi-supervised learning algorithms that consider texts represented in a vector space model usually obtain unsatisfactory classification performances and are surpassed by semi-supervised learning algorithms that consider texts represented in a network. Nevertheless, despite the classification performances, effective approaches based on networks are generated through the similarities among documents and the classification of a new document are also based on the computation of similarities. This implies to set parameters and compute similarities to both generation the networks and classification of new documents. This approach is not feasible to generate fast responses and consequently to classify a huge volume of texts. In this article, we propose an approach to induce a classification model through semi-supervised learning considering text collections represented by bipartite heterogeneous networks. Bipartite networks are easily and quickly generated, leading to classification performance equivalent or better than other approaches based on network or vector space model and allows a fast classification of new documents. The results presented in this article demonstrate that the proposed approach is able to (i) speed up semi-supervised learning, (ii) speed up the classification of new documents and (iii) surpass classification performance of other existing inductive semi-supervised learning techniques.

Keywords: Text Classification, Transductive Learning, Graph-based Learning, Label Propagation, Bipartite Heterogeneous Network.

*Corresponding author
Email addresses: rafael.g.rossi@ufms.br (Rafael Geraldeli Rossi), alneu@icmc.usp.br (Alneu de Andrade Lopes), solange@icmc.usp.br (Solange Oliveira Rezende)