稿件编号： 3088

Title：Inductive Model Generation for Text Classification Using a Bipartite Heterogeneous Network

中文题目：使用异构网络的文本分类归纳模型生成研究

Abstract: Algorithms for numeric data classification have been applied for text classification. Usually the vector space model is used to represent text collections. The characteristics of this representation such as sparsity and high dimensionality sometimes impair the quality of general-purpose classifiers. Networks can be used to represent text collections, avoiding the high sparsity and allowing to model relationships among different objects that compose a text collection. Such network-based representations can improve the quality of the classification results. One of the simplest ways to represent textual collections by a network is through a bipartite heterogeneous network, which is composed of objects that represent the documents connected to objects that represent the terms. Heterogeneous bipartite networks do not require computation of similarities or relations among the objects and can be used to model any type of text collection. Due to the advantages of representing text collections through bipartite heterogeneous networks, in this article we present a text classifier which builds a classification model using the structure of a bipartite heterogeneous network. Such an algorithm, referred to as IMBHN (Inductive Model Based on Bipartite

Heterogeneous Network), induces a classification model assigning weights to objects that represent the terms for each class of the text collection. An empirical evaluation using a large amount of text collections from different domains shows that the proposed IMBHN algorithm produces significantly better results than k-NN, C4.5, SVM, and Naive Bayes algorithms.

中文摘要：一些用于数字数据分类的算法被用于文本分类。通常，向量空间模型被用于文本表示。这种文本表示方法的特点如稀疏性和高维度通常会降低通用分类器的性能。为解决上述问题，网络方法被用于文本集合表示以对文本间的关系进行建模。一种最为简单的利用网络表示文本的方法就是利用二步异构网络，它包括分别表示文本和词项的节点。二步异构网络能够对任何类型的文本集合进行建模并且无需计算文本节点之间的相似度。本文提出了一种名为基于归纳模型的二步异构网络文本分类器，它通过分类模型给出词项在文本类别上的权重。大量的在不同领域文本集合上的实验结果表明，基于归纳模型的二步异构网络文本分类算法在文本分类上要比*k*-NN, C4.5, 支持向量机和朴素贝叶斯的分类效果要好。

Keywords: heterogeneous network, text classification, inductive model generation

中文关键词：异构网络，文本分类，归纳模型生成