计算机工程与应用 2006 42 (22期): ISSN: 1002-8331 CN: 11-2127/TP

博士论坛

大规模文本数据库中的短文分类方法

王永恒, 贾焰, 杨树强

湖南长沙国防科技大学计算机学院网络所613室

收稿日期 2006-3-22 修回日期 网络版发布日期 接受日期

摘要 信息技术的飞速发展造成了大量的文本数据累积，其中很大一部分是短文本数据。文本分类技术对于从这些海量短文中自动获取知识具有重要意义。但是由于短文中的关键词出现次数少，而且带标签的训练样本又通常数量很少，现有的一般文本挖掘算法很难得到可接受的准确度。一些基于语义的分类方法获得了较好的准确度但又由于其低效性而无法适用于海量数据。文本提出了一个新颖的短文分类算法。该算法基于文本语义特征图，并使用类似kNN的方法进行分类。实验表明该算法在对海量短文进行分类时，其准确度和性能超过其它的算法。

关键词 文本挖掘,分类,短文,大规模文本数据库

分类号

SHORT DOCUMENTS CLASSIFICATION METHOD IN VERY LARGE TEXT DATABASE

,,

湖南长沙国防科技大学计算机学院网络所613室

Abstract

With the rapid development of information technology, huge data is accumulated. A vast amount of such data appears as short documents. It is very useful to classify such short documents to get knowledge automatically form the data. But most of the current classification algorithms can’ t get acceptable accuracy since key words appear less time in short documents and the labeled training examples are usually very few. Some classification algorithms based on semantic information is more accurate but they are inefficient to be used to process very large document sets. In this paper, we propose a novel classification method based on semantic text features graph and kNN like method. Our experimental study shows that our algorithm is more accurate and efficient than other classification algorithms when classifying large scale short documents.

Key words text mining classification short document very large text database

DOI:

通讯作者 王永恒 tommywang tommywang3465@hotmail.com