<<文本挖掘中若干关键问题研究>>

图书基本信息

书名：<<文本挖掘中若干关键问题研究>>

13位ISBN编号：9787312022807

10位ISBN编号：7312022804

出版时间：2008-12

出版时间：中国科学技术大学出版社

作者：陆旭

页数：117

<<文本挖掘中若干关键问题研究>>

内容概要

本书介绍了文本分类和偏最小二乘回归，提出了基于变量投影重要性指标的文本分类特征选择方法，论述了偏最小二乘Logistic文本分类模型，阐述了CHTC层次文本分类模型的研究工作，本书可供相关领域科研工作者、大学高年级学生和研究生阅读。

<<文本挖掘中若干关键问题研究>>

书籍目录

章节摘录

第2章 文本分类概述 2.5 文档的表示 2.5.1 文档的特征 对文档进行预处理以后，需要根据文本分类模型对文档进行相应的特征表示，从文档的组成来看，它是字符串的集合，一般来说，文档的特征项应该具有以下特点：特征项是能够对文档进行充分表示的语言单位；文档在特征项集合上的分布具有较为明显的统计规律；特征项分离比较容易实现，计算复杂度不太大，在文本分类中，按照文档特征的粒度来划分，常用的特征单位有词、词组、N—Gram（N元）项和概念等，中文有时也把词性作为文档的特征， 1.词 在信息检索领域，词（Word）是使用最为普遍的文档特征，英语、法语和德语等西方语言通常采用空格或标点符号将词隔开，具有天然的分隔符，所以词的获取简单，中文、日文和韩文等东方语言，句子之间有分隔符，但词与词之间没有分隔符，所以需要分词来得到词。

编辑推荐

《文本挖掘中若干关键问题研究》:自动文本分类是将自然文本文件根据内容自动分为预先定义的一个或几个类别的过程，基于统计学习、机器学习的文本分类技术已经成为主流技术，《文本挖掘中若干关键问题研究》对基于统计学习的文本分类及其相关技术进行了研究，为解决文本分类的稀疏性和高维性问题，基于偏最小二乘理论，提出一种新的维数约简算法，从提高文本分类性能和准确性出发，运用偏最小二乘的最新理论成果，提出了一种能较好提取潜在语义的新文本分类模型，对于数量庞大的文档类别，传统的平坦文本分类的性能受到很大的制约，层次文本分类是一种有效的解决方法，由此提出了一种新的层次文本分类模型。

版权说明