<<Web数据挖掘>>

图书基本信息

书名：<<Web数据挖掘>>

13位ISBN编号：9787302298700

10位ISBN编号：730229870X

出版时间：2013-1

出版时间：清华大学出版社

作者：刘兵

页数：434

译者：俞勇

<<Web数据挖掘>>

内容概要

过去几十年里，Web的迅速发展使其成为世界上规模最大的公共数据源。
Web挖掘的目标是从Web超链接、网页内容和使用日志中探寻有用的信息。
《世界著名计算机教材精选：Web数据挖掘（第2版）》旨在阐述Web数据挖掘的概念及其核心算法，使读者获得相对完整的关于Web数据挖掘的算法和技术知识。
本书不仅介绍了搜索、页面爬取和资源探索以及链接分析等传统的Web挖掘主题，而且还介绍了结构化数据的抽取、信息整合、观点挖掘和Web使用挖掘等内容，这些内容在已有书籍中没有提及过，但它们在Web数据挖掘中却占有非常重要的地位。
全书分为两大部分：第一部分包括第2章到第5章，介绍数据挖掘的基础，第二部分包括第6章到第12章，介绍Web相关的挖掘任务。
从本书自第1版出版之后，很多领域已经有了重大的进展。
新版大部分的章节都已经添加了新的材料来反应这些进展，主要的改动在第11章和第12章中，这两章已经被重新撰写并做了重要的扩展。
《世界著名计算机教材精选：Web数据挖掘（第2版）》不仅可作为本科生的教科书，也是在Web数据挖掘和相关领域研读博士学位的研究生的重要参考用书，同时对Web挖掘研究人员和实践人员获取知识、信息、甚至是创新想法也很有帮助。

作者简介

作者：（美国）刘兵 译者：俞勇

<<Web数据挖掘>>

书籍目录

章节摘录

版权页： 插图： 5.2.1 PU学习的应用 由于人们在大多数情况下仅仅对某个特定类别的网页或文本文档感兴趣，所以在网页和文本文档的检索中PU学习问题经常出现。
例如，某些人可能只对与旅游相关的网页（正例网页）有兴趣，这时所有其他网页都可以被看成是反例网页。
下面让我们通过一个具体的例子来看看PU学习应用的真实场景。
例1：我们想要建立一个关于数据挖掘研究的论文库。
首先，我们可以从一些数据挖掘的会议或者期刊上选取一些论文作为初始的论文集。
然后，我们希望从一些在线的关于数据库和人工智能领域的会议和期刊中寻找关于数据挖掘的论文。
在这些领域的会议和期刊论文中都包含有一些数据挖掘的论文。
同样它们也包含很多其他研究领域的论文。
问题就成了怎样从这些会议和期刊论文中抽取数据挖掘的论文，即怎样在没有进行任何反例文档标注的情况下把这些文章分类成数据挖掘论文和非数据挖掘论文。
在实际应用中，正例文档对于那些已经从事某项特定工作很长时间的人来说是很容易得到的，因为他们在工作过程中可能会积累很多相关文档。
即使一开始没有正例文档的话，直接从Web或者其他资源中收集一些正例文档是相对容易的。
这样人们就可以在没有任何反例标注的情况下，通过使用这个初始正例集从其他一些数据来源中去发现相同类别的文档。
PU学习在以下这些情况下十分有用：（1）从多个无标注集中学习：在一些应用中，人们需要从大量文档集中发现正例文档。
例如，我们希望分辨那些销售打印机的网页。
首先，我们可以很容易从某个在线交易网站中获得一些正例网页，如amazon.com。
然后我们希望从其他一些交易网站中找到打印机网页。
为此，我们需要一一爬下每个网站的内容，然后使用PU学习算法从每个网站中抽出打印机网页。
我们不需要对任何网站中的反例网页进行人工标注。
尽管为一个网站标注一些反例网页并不是太难，但是如果要对每个网站都进行标注的话就很困难了。
由于站点S1中的反例网页可能与站点S，中的反例网页十分不同，所以基于S2中的反例网页学习得到的分类器可能不能用于对站点S2的网页分类。
这个原因在于，尽管两个站点都销售打印机，但是它们出售的其他产品可能大相径庭。
因此使用从S1上学习得到的分类器对S2中的网页分类可能会违背机器学习的基本假设：训练数据和测试数据符合相同的数据分布。
从而，我们可能会得到很差的分类精度。

<<Web数据挖掘>>

编辑推荐

《世界著名计算机教材精选:Web数据挖掘(第2版)》不仅可作为本科生的教科书，也是在Web数据挖掘和相关领域研读博士学位的研究生的重要参考用书，同时对Web挖掘研究人员和实践人员获取知识、信息、甚至是创新想法也很有帮助。

版权说明