<<Web知识挖掘>>

图书基本信息

书名：<<Web知识挖掘>>

13位ISBN编号：9787030274991

10位ISBN编号：7030274997

出版时间：2010-6

出版时间：科学出版社

作者：郑庆华

页数：336

<<Web知识挖掘>>

前言

1989年，欧洲核子研究组织（European Organization for Nuclear Research，CERN）的工程师Tim Berners-Lee针对科学家之间文献交流的需求，首次提出了Web概念与应用架构，其核心是通过超链接实现文本文档的共享。
其后，随着超文本标记语言（HTML）、超文本传输协议（HTTP）等技术标准的逐步成熟，以及Mosaic、Narigatot等浏览器的广泛应用，Web在1995年前后进入了快速发展阶段，表现为Inter。net上的Web页面数量与服务器数量呈指数级增长。
2004年以后，Internet上的PIW（pIablicly indexable Web）页面数已达到了10数量级，每天新增页面的数量超过800万，而Web服务器数量的倍增周期仅为23周。
Web已成为一个开放性的、动态的、全球性信息服务中心，以及当前人们获取信息的重要手段。
然而，Web上同样面临着信息社会所共有的“信息爆炸”与“知识贫乏”的矛盾性问题。
如何从这些海量的Web数据中发现有用的知识或者模式，成为人们亟待解决的问题。
传统的数据挖掘技术主要针对结构化的数据对象，还很难适用于具有异构性、半结构化特性以及高度动态性等特点的Web数据。
为此，Etzioni于1996年提出了“Web挖掘”的概念。
Web挖掘是数据挖掘、机器学习、数据库、自然语言处理、Web／Internet等多种信息技术相互渗透与融合的必然结果，旨在研究如何从Web文档与服务中抽取有价值的知识或隐含信息。
近年来，Web挖掘这个研究领域得到了国内外学者越来越多的关注，人们以文本分类、信息抽取、检索结构排序、用户访问模式发现等应用为目标，在Web挖掘的三个子领域——结构挖掘、内容挖掘、日志挖掘方面从事了大量的研究工作，在理论、方法与应用方面取得了一系列研究成果。

<<Web知识挖掘>>

内容概要

本书是一部关于Web知识挖掘的比较系统、完整，且理论和实践相结合的著作，共含7章：第1章与第2章是Web知识挖掘概论，其中，第1章总体上对Web知识挖掘的现状、概念、典型方法、应用领域以及面临的挑战进行综述性说明；第2章介绍了Web知识挖掘的预备知识、分类体系、基本流程等内容。第3～6章是Web知识挖掘的理论与方法，分别论述了Web爬取、Web结构挖掘、内容挖掘、日志挖掘相关理论与方法，并系统总结了我们自己在元数据、概念、知识元等多个层次上的知识获取以及个性化知识服务等方面的工作。
第7章是Web知识挖掘的实践与应用实例，以实例对Web结构挖掘、日志挖掘及内容挖掘的应用进行了说明。
　本书不仅系统地介绍了Web知识挖掘领域的基础理论与方法，也阐述了我们在该领域的创新性工作，因而适合不同类型与层次的研究人员及学生。
　本书可作为信息领域的科研与工程技术人员的参考书，也可作为计算机与相关专业的研究生和高年级本科生的教材或辅导书目。

<<Web知识挖掘>>

书籍目录

章节摘录

插图：Web挖掘是从数据挖掘发展而来的，但与传统的数据挖掘相比有许多独特之处。
数据挖掘，又称为面向数据库的知识发现（knowledge discovery in database，KDD），就是从大量数据中获取新颖的、潜在有用的模式的过程。
数据挖掘的对象是来自关系型数据库或XML数据库中的结构化数据。
而Web挖掘的对象包括网页、图像、声音、视频、网页之间的链接以及网站用户的日志数据。
除了日志数据外，其他类型数据具有海量、异构、非结构化等特性，传统的数据挖掘技术还很难处理这类数据。
因此，必须在Web挖掘领域中，研究专门针对Web数据特点的算法与方法。
在信息检索中，用户以关键词组合表达检索需求，通过关键词匹配的方式从特定文档集中返回与检索需求相关的文档。
信息检索包括文档的建模、分类、索引、结果排序与可视化Web等流程，Web挖掘技术一般用于其中的分类、索引以及结果排序，从这个角度来说，Web挖掘是信息检索过程的重要组成部分（Kosala etal，2000）。
另一方面，信息检索的结果往往也是Web挖掘的对象，如在HITS算法中，因而信息检索也可作为Web挖掘的组成部分。
信息抽取指从给定的文档中抽取特定类别的信息，例如，从一篇文档中抽取标题、作者等元数据信息。
由于Web站点的异构性，大多数信息抽取都是对针对特定网站，一些抽取方法能够自动或半自动地建立抽取模式（Kushmerick，1999），对于这类信息抽取，Web挖掘可以看做信息抽取的一个过程。
此外，在Web挖掘中，利用信息抽取可以建立文档的压缩版本以提高挖掘效率，从这个角度来说，信息抽取可以作为Web挖掘的预处理过程。

编辑推荐

《Web知识挖掘:理论、方法与应用》由科学出版社出版。

版权说明