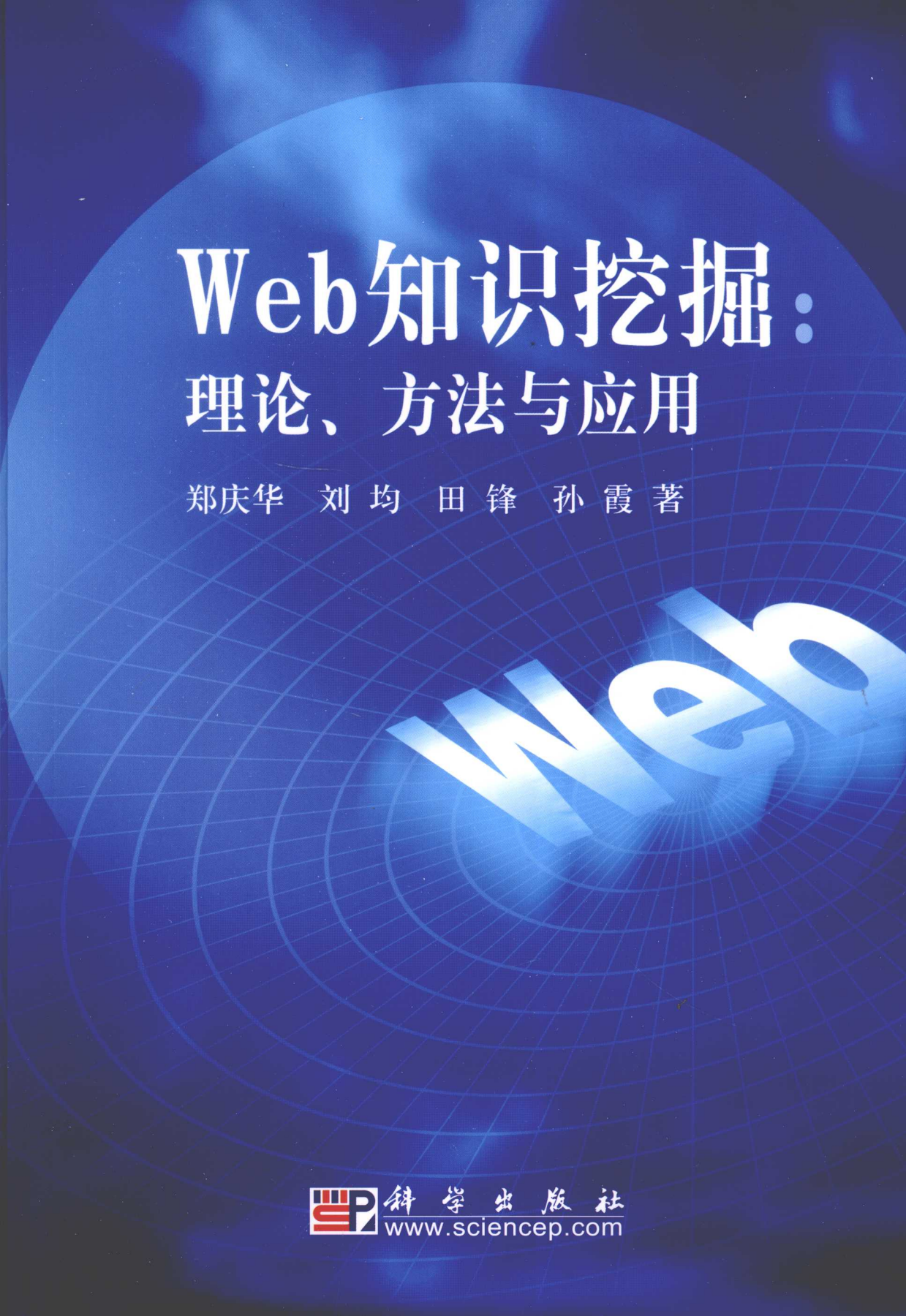
Web知识挖掘：理论、方法与应用

郑庆华　刘均　田锋　孙霞　著



科学出版社
www.sciencep.com

Web 知识挖掘：理论、方法与应用

郑庆华　刘　均　田　锋　孙　霞　著

科学出版社
北京

内 容 简 介

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。

图书在版编目(CIP)数据

Web知识挖掘：理论、方法与应用/郑庆华等著．—北京：科学出版社，2010.6

ISBN 978-7-03-027499-1

Ⅰ.①W… Ⅱ.①郑… Ⅲ.①计算机网络-情报检索 Ⅳ.①G354.4

中国版本图书馆CIP数据核字（2010）第083171号

责任编辑：刘宝莉 任 静 / 责任校对：陈玉凤
责任印制：赵 博 / 封面设计：耕者设计工作室

科学出版社 出版
北京东黄城根北街16号
邮政编码：100717
http://www.sciencep.com

西源印刷厂 印刷

科学出版社发行 各地新华书店经销

*

2010年6月第 一 版 开本：B5（720×1000）
2010年6月第一次印刷 印张：21 3/4
印数：1—3 000 字数：403 000

定价：50.00元

（如有印装质量问题，我社负责调换）

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、Navigator 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 Internet 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（publicly indexable Web）页面数已达到了 10^{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。

作者多年来在国家自然科学基金、863 等国家级课题的支持下，以 E-learning 为应用背景，在 Web 挖掘领域，特别是内容挖掘与日志挖掘方面，开展了大量深入的研究工作；并且从 2003 年起，连续 7 年为西安交通大学电信学院的博士生与硕士生讲授 Web 挖掘课程。本书正是基于这些工作而撰写的。我们希望本书不仅为 Web 挖掘领域的发展做出贡献，更希望其成为我们与国内外同行交换学术见解的桥梁。

本书内容

本书共有 7 章，按 Web 知识挖掘概论、理论与方法、实践与应用三个部分进行

组织：

(1) Web 知识挖掘概论。包括第 1 章与第 2 章，其中，第 1 章总体上对 Web 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。第 2 章介绍了 Web 挖掘的预备知识、分类体系、基本流程以及该领域的重要文献、国际会议等内容。

(2) Web 知识挖掘的理论与方法。包括第 3～6 章，其中，第 3 章介绍了 Web 爬取的基本知识，并重点讨论了隐含 Web 爬取与面向主题的 Web 爬取两个热点研究问题。第 4 章介绍了经典的 PageRank 算法与 HITS 算法，并讨论了基于超链接的 Web 宏观结构特性分析问题。第 5 章对面向 Web 的页面分类与聚类、多媒体数据挖掘相关理论与方法进行了论述，重点阐述了我们在元数据抽取、本体学习、知识元及其关联抽取等方面的研究工作。第 6 章介绍了 Web 日志预处理方法，并结合数据挖掘技术，重点论述了用户行为模式挖掘、个性挖掘、兴趣感知等理论与方法。

(3) Web 知识挖掘的实践与应用实例。第 7 章首先通过“Web 站点的自适应重构”、“面向网络学习的学习者个性挖掘与个性化学习”、“海量 Web 资源中的知识处理与服务”三个实例对 Web 结构挖掘、日志挖掘及内容挖掘的应用进行了说明；最后介绍了我们自己研制的 Web 挖掘实验平台。

本书特色

与国外同类书籍相比，本书具有以下特色：

(1) 既注重体现自己的研究特色，也注重 Web 知识挖掘领域知识的系统性。一方面，结合作者在该领域国际期刊与国际会议上发表的数十篇学术论文，将元数据抽取、本体学习、知识元及其关联抽取以及个性挖掘等方面的创新性研究成果融入到 Web 知识挖掘体系中；另一方面，本书也系统地论述 Web 知识挖掘三个方向——内容挖掘、结构挖掘以及日志挖掘的基本理论与方法。

(2) 既注重 Web 知识挖掘领域的基础理论、方法，同时也注重方法的应用。结合作者所承担的国家课题，提炼出实际案例与真实数据，并据此对这些方法的应用进行说明。

适用读者

本书不仅阐述了作者在 Web 内容挖掘与日志挖掘方面的创新性研究工作，而且也系统地介绍了 Web 挖掘领域的基础理论与方法，因而适合于不同类型与层次的研究人员及学生。本书既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。

本书由西安交通大学计算机系郑庆华、刘均、田锋、孙霞撰写。其中，郑庆华负

责本书的第1章、第2章、第5章，刘均负责第3章、第4章、第7章，田锋负责第6章；孙霞讲师(西北大学计算机系)、吴茜媛讲师以及研究生常晓、邓万宇、王伟、刘广东、王世斌、田振华、董博、杜瑾、丁娇、蒋路、许凌志、吴朝晖、林鹏、王艳烨、周正、骞雅楠、沙莎、刘子奇等也参与了本书的撰写工作。

由于Web挖掘是一个新兴的、多学科交叉的研究领域，涉及的范围非常广泛，加之我们自身学识有限，书中疏漏之处在所难免，敬请专家和读者批评指正。

目　录

第 1 章　Web 挖掘概述

本章对 Web 挖掘的产生背景、基本概念、主要应用、面临的挑战，以及主要研究方向等问题进行了描述：①介绍了 Web 的发展历史、趋势以及 Web 信息爆炸的问题；②给出了与 Web 挖掘相关的若干基本概念，阐明了 Web 挖掘与数据挖掘、信息抽取、信息检索之间的区别与联系；③给出了 Web 挖掘的若干典型应用；④从 Web 数据自身的复杂性与当前 Web 信息检索的局限性这一矛盾出发，阐述了 Web 挖掘所面临的理论与技术挑战；⑤介绍了 Web 挖掘的主要研究方向与热点问题。

1.1　Web 发展历史与现状

1.1.1　Web 技术发展

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，在所负责的 Enquire（enquire within upon everything）项目基础上，首次提出了 Web 概念与应用架构，其核心是利用超文本标记语言（hypertext markup language，HTML）实现信息与信息的连接；利用统一资源定位（uniform resource locator，URL）技术实现信息的精确定位；利用超文本传输协议（hypertext transfer protocol，HTTP）实现分布式的信息共享。1990 年 11 月，第一个 Web 服务器 nxoc01. cern. ch 开始运行。1991 年，CERN 正式发布了 HTML、HTTP 等 Web 技术标准。目前，与 Web 相关的技术标准都由 W3C 组织（World Wide Web Consortium）管理和维护。

Web 应用架构是一种 Client/Server 架构，相应地，Web 技术可被分为客户端和服务端两大类。以下简单说明两类技术的发展现状。

1. 客户端技术

Web 客户端（浏览器）技术是指集成于 Web 浏览器的技术，涉及 HTML 语言、Java 语言、级联样式表（cascading style sheets，CSS）、DHTML（dynamic HTML）以及浏览器插件等。总体上，Web 客户端技术是由静态向动态逐渐发展起来的。

HTML 语言的历史最早可以追溯到 20 世纪 40 年代。1945 年，Bush 首先提

出了超链接的思想；1963～1965 年，Nelson 最先提出了超文本(hypertext)与超媒体(hypermedia)的术语。1969 年，Goldfarb 发明了描述超文本信息的 GML(generalized markup language)语言，在 ANSI 等组织的努力下，GML 进一步发展成为标准通用标记语言(standard generalized markup language，SGML)。HTML 语言是 SGML 的简化和完善，适合超文本信息的传递和解析，但是只能展示静态的文本与图像，满足不了人们对交互性的需求，Java 语言、CSS、DHTML 以及各种浏览器插件技术改变了这一现状。

Java 语言的平台无关性，适合浏览器中动态应用的开发。1996 年，Netscape 2.0 与 IE 3.0 开始支持 JavaApplets 和 JavaScript。1996 年，W3C 提出了 CSS 的建议标准，CSS 提高了对信息展现的控制能力。1997 年，Microsoft 集成动态 HTML标记、CSS 和动态对象模型(DHTML object model)，形成了一套完整、实用、高效的客户端开发技术体系 DHTML。DHTML 无需启动 Java 虚拟机或其他脚本环境，就可获得更好的展现效果和更高的执行效率(王咏刚，2004)。

2. 服务器端技术

与客户端技术的发展过程类似，Web 服务端技术也是由静态向动态逐渐发展起来的。

最早的 Web 服务器简单地响应浏览器发来的 HTTP 请求，并将服务器上的 HTML 文件返回给浏览器。CGI(common gateway interface)技术是第一种动态生成 HTML 页面的技术。早期的 CGI 程序大多是编译后的可执行程序，编程语言是 C、C++、Pascal 等。为了简化 CGI 程序的修改、编译和发布过程，人们开始使用 Perl、Python 等脚本语言。

1994 年，Lerdorf 发明了 PHP(personal home page tools)语言。PHP 语言将 HTML 代码和 PHP 指令合成为完整的服务端动态页面。1997～1998 年，Servlet 技术与 JSP 技术诞生，让开发者同时拥有了类似 CGI 程序的集中处理功能和类似 PHP 的 HTML 嵌入功能(王咏刚，2004)。

1.1.2 Web 上的信息爆炸

随着 HTML、HTTP 等 Web 技术的逐步成熟，以及 Mosaic、Navigator、IE 等浏览器的广泛应用，Internet 上的 Web 网站数目与页面数目都呈指数级的速度增长。Web 网站数目从 1990 年的 1 个发展到 2006 年超过 10^8 个(Berners-Lee，2007)，倍增周期仅有 23 周，图 1-1 是 Web 网站从 1990～2006 年增长情况(Zakon，2006)。Web 页面数目则平均每 6 个月翻一番，2004 年以后，Internet 上的页面已达到了 10^{10} 数量级，每天新增页面的数目超过 800 万。

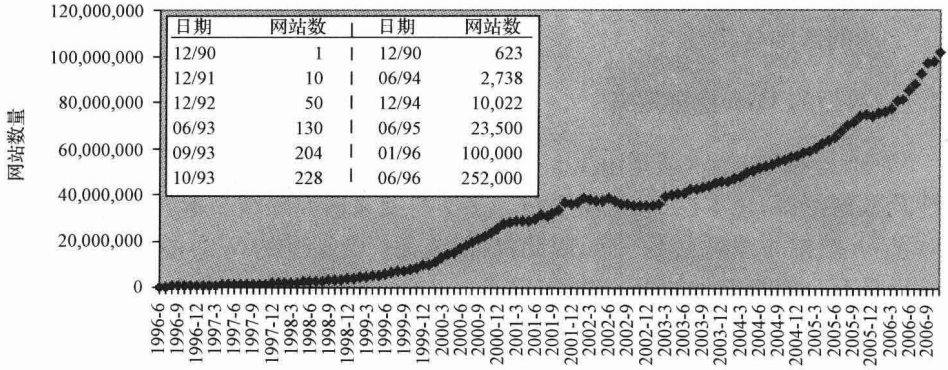


图 1-1　Web 网站的增长（Zakon，2006）

然而，上述页面仅仅是能被搜索引擎搜索到的网页，称为 PIW(publicly indexable Web)或 Surface Web 页面。绝大多数 Web 页面是用户以特定的查询接口查询 Web 数据库过程中动态生成的，这类页面无法被传统的搜索引擎爬取到，被称为 Deep Web(或 Hidden Web)页面。据 Brightplanet 公司的研究人员分析，Deep Web 中的页面数量是 PIW 页面 400～550 倍，约占整个 Web 页面 99.7%，并且页面质量也远远高于 PIW 页面(Bergman，2001)。

Web 的海量信息一方面为人们提供了获取信息的源泉，另一方面，也为人们高效地获取有价值的信息带来了挑战，表现为：①Web 信息的总量虽然很大，但对于某一个特定用户，他所感兴趣的 Web 信息却相对很少，即“99%的 Web 信息对于 99%的 Web 用户是没有用处的”，即所谓的“丰度问题”(Dunham，2002；Kleinberg，1998)；②作为获取 Web 信息的重要手段，当前各种主流搜索引擎一般都采用关键词或者是关键词的逻辑组合作为检索条件，这种检索技术不支持语义检索，很难明确地表达用户的检索意图；③现有 Web 服务给用户提供的是无差别的、“千人一面”的服务模式，由于不同用户的知识背景以及所感兴趣的领域不同，对 Web 服务的需求也存在很大差异。

Web 技术的快速发展与 Web 信息的迅猛增长使得人们不再满足于 Web 仅作为一个信息共享与发布的平台，如何通过 Web 文档与服务中的挖掘获取有价值的知识或隐含信息，并以此提供智能化、个性化、语义化的信息服务已成为人们的迫切需求。Web 挖掘(Web mining)就是在这样的背景下产生的。

1.2　Web 挖掘的概念

本节首先给出几个常见的 Web 挖掘定义，通过对 Web 挖掘对象、目标以及技术手段的分析，给出我们自己的 Web 挖掘定义；其次，阐述了 Web 挖掘与数据挖

掘(data mining)、信息检索(information retrieval)、信息抽取(information extraction)之间的区别与联系。

1.2.1 典型的 Web 挖掘定义

“Web 挖掘”这个术语是由 Etzioni 于 1996 年提出来的，并逐步发展为一个新的涉及数据挖掘、文本挖掘、机器学习等多学科交叉研究领域。不同研究领域的学者对 Web 挖掘的理解也不一致，因此，目前尚无广泛接受的 Web 挖掘定义。以下是给出具有一定影响力的 Web 挖掘定义。

Etzioni 将 Web 挖掘定义为“利用数据挖掘技术自动从 Web 文档与服务中发现或抽取信息”(Etzioni，1996)。

Srivastava 借鉴数据挖掘的定义(Fayyad et al.，1996)，将 Web 挖掘定义为“从 Web 文档和 Web 活动中抽取感兴趣的潜在的有用模式和隐藏的信息”(Srivastava et al.，2000)。

在维基百科上，Web 挖掘被定义为“利用数据挖掘技术从 Web 中发现模式”(Wikipedia，2007)。

上述定义从挖掘对象、目标以及技术手段三方面对 Web 挖掘进行了简单描述。Web 挖掘经过了十余年的发展，其挖掘对象、目标以及技术手段逐步得到具体化。

挖掘对象主要包括：①隐藏在半结构化数据中的模式和数据实体，包括半结构化或非结构化的文本数据、图像数据、视音频数据及其元数据；②描述内容格式规定与组织结构的数据，如超链接关系；③用户访问页面内容时的记录数据，如 Web 服务器日志。

挖掘目标为规则、模式、领域知识、特定实体及其关联、限制条件等。

挖掘的技术手段：Web 挖掘涉及数据挖掘、文本挖掘、机器学习、数据库、Web/Internet、等多个领域的理论与方法。数据挖掘只是 Web 挖掘的技术手段之一，并且主要应用于 Web 日志挖掘中。

在对 Web 挖掘的对象、目标以及技术手段分析的基础上，我们提出了自己的 Web 挖掘定义：“利用数据挖掘、文本挖掘、机器学习等技术从 Web 页面数据、日志数据、超链接关系中发现感兴趣的、潜在的、有用的规则、模式、领域知识等”。

1.2.2 Web 挖掘与数据挖掘、信息检索、信息抽取的区别

Web 挖掘是一个多学科交叉的领域，挖掘过程中，频繁使用了数据挖掘、信息检索、信息抽取等多种技术。Web 挖掘与这些技术领域既有一定的联系，又具有显著的区别。

1. Web 挖掘与数据挖掘

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。

2. Web 挖掘与信息检索

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 et al.，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。

3. Web 挖掘与信息抽取

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。

1.3　Web 挖掘面临的挑战

Web 挖掘的对象是 Web 页面数据、页面之间的超链接数据以及用户访问的日志数据(Srivastava et al.，2000)，由于这类数据自身的复杂性以及在获取手段方面的局限性，导致 Web 挖掘与传统的数据挖掘相比，面临着一些新的挑战。

1.3.1　Web 数据的高度复杂性

Web 数据的复杂性体现在数据的异构性、半结构化特性、动态性以及存在噪

声数据等多个方面。

(1) 异构性。Web 数据的异构性表现为两个方面：一方面，作为挖掘对象的页面数据(包括文本、多媒体等)、超链接数据以及日志数据本身是异构的；另一方面，作为数据源的每个 Web 站点，其信息组织方式也是异构的。针对 Web 数据的异构性，需要研究的问题主要包括各种异构数据对应的挖掘方法、基于统一的视图异构数据源的集成。

(2) 半结构化特性。数据模式用于定义数据的结构、属性、联系以及约束。根据数据及其模式的独立性，数据可分为结构化数据、半结构化数据以及非结构化数据。同传统关系型数据库中的数据一样，对于 Web 日志数据，其数据与模式是完全独立的，因此，这类数据是结构化数据。而对于 Web 页面数据，模式信息与数据值混合在一起，这种自描述的数据就是半结构化数据。针对 Web 页面数据的半结构化特性，需要对页面数据的建模、集成与检索等问题开展研究。

(3) 动态性。Web 数据具有高度的动态性(Han et al., 2002)。这种动态性不仅表现在 Web 数据量指数级的增长，而且也表现在已有数据的频繁更新。被动态改变的数据既包括页面内容，如新闻、公告、股票等，同时也包括超链接数据与用户访问的日志数据。Web 数据的动态性要求 Web 挖掘方法能够发现数据在动态改变过程中的时序规律，同时还应具有较高的效率，满足时效性需求。

(4) 存在噪声数据。Web 数据中还存在大量的噪声数据(noisy data)，这些数据可能会干扰挖掘结果的质量，对于一些噪声敏感的挖掘算法，这些数据可能会导致挖掘结果出现大的偏差。噪声数据主要有两种来源(Liu, 2006)：一是页面中与挖掘应用无关的信息，如广告、版权声明等；二是质量低下的页面信息，由于缺少信息的质量控制机制，人们可以在 Web 上发布任意信息，这些信息的质量与正确性都无法保证。对于细粒度的 Web 挖掘，需要研究如何识别并去除噪声数据，以保证挖掘质量。

1.3.2 Web 数据检索的局限性

搜索引擎是获取 Web 数据的重要手段，然而由于 Web 数据的复杂性以及网络爬虫(Spider)、搜索引擎自身存在的缺陷，导致当前搜索引擎在 Web 数据检索方面还存在以下问题：

(1) 丰度问题(abundance problem)。丰度问题是由美国 Cornell 大学 Kleinberg 教授提出的(Kleinberg, 1998)，表现为 Web 信息的总量虽然很大，但对于某一个特定用户，他所感兴趣的 Web 信息却相对很少，即“99%的 Web 信息对于 99%的 Web 用户是没有用处的”(Dunham, 2002)。例如，在 Google 中以“data mining trend”作为检索条件，检索关于“data mining”研究趋势的文献，Google 共返回 422,000 个结果，但第一页(前 10 个结果)中没有既符合检索条件且具有一定