21世纪高等院校电子商务教育系列教材

基于语义的Web数据挖掘

马 刚 主编

Semantic Web Data Mining

东北财经大学出版社
Dongbei University of Finance & Economics Press

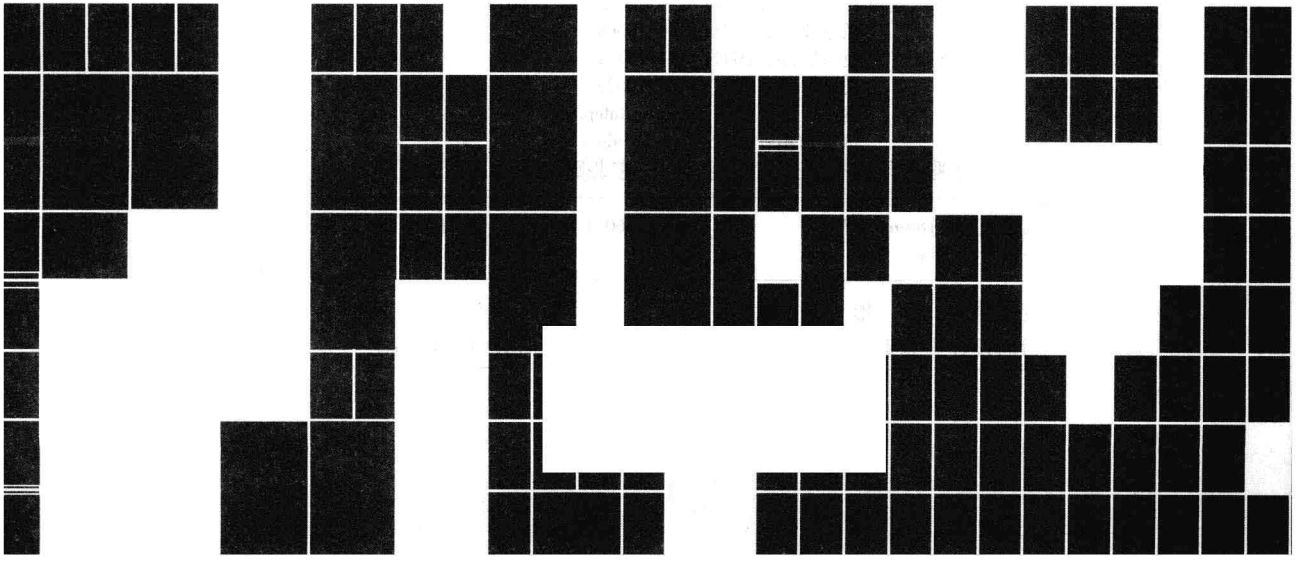
21世纪高等院校电子商务教育系列教材

基于语义的Web数据挖掘

马 刚 主编

Semantic Web Data Mining

东北财经大学出版社
Dongbei University of Finance & Economics Press



图书在版编目（CIP）数据

基于语义的 Web 数据挖掘 / 马刚主编 . —大连 ：东北财经大学出版社，2014. 1
（21 世纪高等院校电子商务教育系列教材）
ISBN 978-7-5654-1375-9

Ⅰ. 基… Ⅱ. 马… Ⅲ. 数据采集-高等学校-教材
Ⅳ. TP311. 13

中国版本图书馆 CIP 数据核字（2013）第 271102 号

东北财经大学出版社出版
（大连市黑石礁尖山街 217 号 邮政编码 116025）
教学支持：（0411）84710309
营 销 部：（0411）84710711
总 编 室：（0411）84710523
网 址：http：//www. dufep. cn
读者信箱：dufep @ dufe. edu. cn
大连美跃彩色印刷有限公司印刷 东北财经大学出版社发行

幅面尺寸：186mm×230mm 字数：560 千字 印张：26
2014 年 1 月第 1 版 2014 年 1 月第 1 次印刷

责任编辑：李 彬 王 斌 责任校对：百 果
封面设计：冀贵收 版式设计：钟福建

ISBN 978-7-5654-1375-9
定价：42. 00 元

前言

随着互联网的发展，近年来关于 Web 数据挖掘的研究方兴未艾，加之多年从事数据挖掘方面的教学工作，笔者一直希望撰写一本这方面的书，以飨读者，其目的就是作为想了解和进入 Web 数据挖掘研究和实践领域的工作者的“他山之石”，帮助他们把握本领域的全貌，掌握研究和解决 Web 数据挖掘问题的切入点。

伴随互联网的应用，社区、论坛、微博上留下了浩瀚的数据信息，这些数据蕴藏着巨大的商机和社会价值。与此同时，Web 上信息爆炸与知识贫乏的矛盾依然存在，如何在海量的 Web 数据中发现知识，并用于社会实践，仍然是亟待解决的问题。

Web 数据挖掘的发展经历了三个阶段。第一阶段，也可以说是传统的 Web 数据挖掘阶段，Web 数据挖掘的对象主要集中于网站内的信息，分为内容挖掘、结构挖掘和使用挖掘。第二阶段，随着互联网的发展，人们开始关注网络浏览者留下的信息，从中发现商机，这就有一个对文本的自然语言理解的问题。特别是中文不同于英文，英文是以词为语言单位的，而中文是以句子为语言单位的。所以，自然语言理解就成为 Web 数据挖掘的一个重要课题。其中，语言的机器自动理解要经过数据抓取，文本的初始化处理、分词，句法分析，词的极性判断，句子的极性判断，文本的极性判断等步骤才能完成，完成这个过程是有困难的。于是，为了改善当代万维网信息不利于计算机自动处理的现状，万维网的创始人 Tim Berners Lee 于 1998 年提出了有关下一代万维网的构想——语义万维网，即

目录

第1章

Web 数据挖掘概述

学习目标

了解 Web 数据挖掘的产生、概念及特点；掌握 Web 数据挖掘与传统数据挖掘、信息检索和信息抽取的区别；掌握 Web 数据挖掘的三种分类及其比较；了解 DSS 的基本知识，弄清 DSS 与 BI 的区别；了解 Web 数据挖掘的应用、面临的挑战以及其发展趋势。

随着网络和通信技术的发展，互联网已经成为全球最大的信息服务平台，给人类的活动带来了巨大的变革。在当今时代，信息显得尤为重要，互联网上的信息资源空前丰富，但同时人们也面临着信息过量的问题。互联网上的信息每天都在以惊人的速度增长，如何在海量的信息中寻找用户感兴趣和需要的信息资源已经成为迫切需要解决的问题。Web 数据挖掘正是由此应运而生的，它可以帮助用户挖掘隐藏在大量信息背后的知识，满足用户的需求，目前，已经成为数据挖掘领域中新兴的研究热点。

1.1 Web 数据挖掘基础

数据挖掘技术是人们长期对数据库技术进行研究和开发的结果。从起初将各种商业数据存储在计算机的数据库中，到后来可以对数据库进行查询和访问，甚至是即时遍历。但是，随着人们积累的数据越来越多，如何从海量的数据中找到内在的规律，获取有用的信息，挖掘这些数据背后隐藏的重要信息已经成为当前高科技领域研究的热点。目前，数据挖掘使数据库技术进入了一个更高级的阶段，它不仅能对过去的数据进行查询和遍历，并且能够找出过去数据之间的潜在联系，从而促进信息的传递。

1.1.1 Web 数据挖掘的产生

随着互联网的快速发展，特别是 Web 技术的发展，使得互联网上出现了大量的 Web 站点，每个站点就是一个数据源，包括用户浏览记录、交易记录、日志文件、网络页面信息以及超链接信息等。因此，Web 上的数据信息正在以惊人的速度增长。但是，在这越来越多的信息当中只有其中的一小部分是我们所关注的。所以如何在海量的、动态的互联网信息数据中获取有用的知识成为数据挖掘领域一个新的挑战，Web 数据挖掘正是在这一背景下产生的。

信息检索技术，如搜索引擎，可以帮助人们尽快地找到所需要的信息，但是目前多数搜索引擎都存在两方面的缺陷，即查全率低和查准率低。也就是说，用户的一个查询请求往往会检索出一个庞大的结果集，而用户所需要的信息却只是其中的很小一部分。而且，利用搜索引擎也不能检索出 Web 中的所有与检索相关的 Web 页面。但是 Web 数据挖掘能够挖掘隐藏在信息背后的知识，能够提供满足用户需求的信息。因此，Web 数据挖掘的发展变得尤为必要。

另外，随着互联网的普及，电子商务也在蓬勃发展，开展电子商务的企业面临着极大挑战，即如何对用户的注册信息、浏览信息、历史购买记录等数据信息进行有效的组织利用，从而了解用户的兴趣爱好、行为模式等，以优化网站结构、发掘潜在用户、为用户提供个性化服务等。这些挑战也在推动着 Web 数据挖掘的发展。

1.1.2　Web 数据挖掘的概念及特点

1. Web 数据挖掘的概念

Oren Etioni 在 1996 年首次提出 Web 数据挖掘这一概念，他认为 Web 数据挖掘是运用数据挖掘技术从 Web 文档和服务中自动地发现和抽取信息。一般情况下，“因特网的数据挖掘”、“Web 数据发现”、“网络信息挖掘”、“Web 信息挖掘”等也可以被认为是 Web 数据挖掘的同义词。Web 数据挖掘是一项综合技术，是数据挖掘技术在 Web 领域中的应用，并与 Web 技术相结合的产物，涉及 Web 技术、数据挖掘、人工智能以及统计学等多个领域。

不同领域的学者对 Web 数据挖掘的理解也不一致，因此，Web 数据挖掘目前没有统一的定义，以下给出一些具有影响力的 Web 数据挖掘的定义。

Srivastava 将 Web 数据挖掘定义为“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，2000）。在维基百科上，Web 数据挖掘被定义为“利用数据挖掘技术从 Web 中发现模式”（Wikipedia，2007）。

本书采用下面更为一般的定义。

【定义 1-1】

Web 数据挖掘是指从大量 Web 文档的集合 C 中发现隐含的模式 P，如果将 C 看做输入，将 P 看做输出，那么，Web 数据挖掘的过程就是从输入到输出的一个映射：C→P。

从传统数据挖掘的概念出发，可以将 Web 数据挖掘理解为，Web 数据挖掘[①]是从大量非结构化、异构的 Web 信息资源中发现有效的、潜在可用的及最终可以理解的知识（包括概念、模式规则、规律、约束及可视化等形式）的非平凡过程。

2. Web 数据挖掘的特点[②]

Web 数据挖掘是一种特殊的数据挖掘，它是在传统的数据挖掘技术的基础上与现代统计分析、人工智能等技术结合产生的。虽然 Web 数据挖掘技术是由传统数据挖掘技术发展而来的，但是它们还是有很多不同之处。

（1）Web 数据挖掘对象的海量性与动态性

Web 是一个不断变化的、动态更新的系统，Web 上的数据信息也是不断更新的，并且，Web 上的数据信息正在以惊人的速度增长。因此，Web 数据挖掘面对的数据源具有很强的动态性。而传统数据挖掘面对的数据源则是静态的。

（2）半结构化的数据结构

传统的数据挖掘以数据库为基础，是对结构化的数据进行加工、分析和模式挖掘。Web 上的数据与传统的数据库中的数据不同，传统的数据库都有一定的数据模型，可以根据模型来具体描述特定的数据，而 Web 上的数据非常复杂，既有数值型数据，也有布尔型数据，

① 周涛，李军，陈惠玲. Web 数据挖掘技术研究［J］. 汉中师范学院报，2004（3）：86-90.
② 王楠. XML 在 Web 数据挖掘中的应用［J］. 科技创新导论，2009（7）：11.

还有描述性数据以及 Web 特有的数据（如 IP 地址），同时还有图像、音频、多媒体等没有特定模型描述的数据，每一站点的数据都各自独立设计，并且数据本身具有自述性和动态可变性，因而 Web 上的数据具有一定的结构性，但因自述层次的存在，从而是一种非完全结构化的数据，也被称为半结构化数据。半结构化是 Web 上数据的最大特点。因此，挖掘对象为结构化数据的传统数据挖掘方法已经不能实现有效挖掘，所以，必须对其进行补充和扩展，才能进行有效的数据挖掘。

（3）异构的数据库环境

Web 上每个站点就是一个数据源，而每个站点的信息组织方式通常不一样，因此每个数据源都是异构的，这样就构成了一个巨大的异构数据库环境。如果想要对这些数据进行数据挖掘，必须先要研究站点之间异构数据的集成问题，只有将这些站点的数据都集成起来，提供给用户一个统一的视图，才有可能从巨大的数据资源中获取所需的东西。

（4）挖掘目的的模糊性

Web 上有成千上万的用户，而且这些用户群还在不断地扩展。同时，又因为每个用户的背景、使用挖掘的目的和兴趣都不同，大多数用户对自己的挖掘主题和应用只有一个肤浅的认识和了解，并不能提出一个明确的目标。所以 Web 挖掘目的是模糊的、不明确的。

（5）Web 数据信息具有分布性、多维性和混沌性

有用信息的获得变得日益困难，绝大多数用户在 Internet 上查询或浏览信息时，往往是通过一些知名的门户网站提供的搜索引擎来实现的，但是无论这个引擎采用什么样的搜索技术，它总是存在着以下两个问题：查准率低和查全率低。所谓查准率低是指当用户输入关键字检索信息时，返回的结果动辄成千上万条，真正有用的也许只有几条，相当一部分检索结果是与检索内容无关的信息，有些甚至是死链接。所谓查全率低是指搜索引擎返回的结果并不能穷举 Web 中所有与检索内容相关的 Web 页，也许返回的结果仅仅是相关内容中很少的一部分。

3. Web 数据挖掘与信息检索、信息抽取的区别[①]

（1）Web 数据挖掘与信息检索

Web 数据挖掘与 Web 信息检索有一些相似的地方，面临的对象都是 Web 文档，都是从大量的 Web 文档中发现资源，人们往往将它们等同起来，但它们是有区别的。

【定义 1-2】

Web 信息检索，是指从大量 Web 文档的集合 C 中找到与给定的查询请求 q 相关的、恰当数目的文档子集 S，Web 信息检索的过程也对应于一个映射：

$F:(C, q) \rightarrow S$

从 20 世纪 60 年代以来，信息检索领域在索引模型、文档内容表示、匹配策略等方面取得了许多研究成果，这些成果被成功地应用在 Web 上，产生了搜索引擎，例如 Google、

① 郑泠. Web 数据挖掘技术应用［J］. 科技经济市场，2006（12）：302-303.

Baidu、Yahoo 等。搜索引擎的基本原理是利用一种被称为“SPIDER”或“ROBOT”的软件工具在 Web 上搜索，将找到的信息编入自己的数据库中，用户检索时直接输入关键词，搜索引擎根据一定的规则将检索词与其数据库中的信息进行匹配，从而生成结果清单。搜索引擎工作的一般流程包括：使用 SPIDER 或 ROBOT 搜集 Web 文档、对文档集合建立倒排索引、分析用户的查询请求、匹配文档与查询请求以计算二者之间的相似度、对查询结果进行排序以及把搜索到的相关信息反馈给用户，具体的流程我们会在后面的章节中讲到。

尽管 Web 信息检索与 Web 挖掘有相似的地方，但它们是两种不同的技术，其区别[①]主要表现在以下几个方面：

①方法论不同。信息检索是目标驱动的，用户根据自己的需要明确提出查询要求，检索系统根据用户的查询要求找到相匹配的信息；而 Web 挖掘独立于用户的信息需求，结果也是用户所无法预知的。

②着眼点不同。信息检索着重于文档中显式存储的字词和链接，Web 挖掘则试图挖掘隐含在 Web 文档中的内容、结构及规律。

③目的不同。Web 信息检索是从大量 Web 文档中找到满足用户查询请求的文档子集，目的是帮助用户查询资源。Web 信息检索通常不能发现隐藏在数据背后的联系，而 Web 挖掘的目的就是从大量 Web 文档和活动中发现潜在的、有用的知识或模式以供决策支持。

④评价方法不同。信息检索使用查准率和查全率来评价其性能，要求返回尽可能多的相关文档，同时不相关的文档尽可能少。查全率是指信息检索系统检索到的相关文档占被检索文档集中所有相关文档的比率，查准率是指信息检索系统检索到的所有文档中相关文档所占的比率。Web 挖掘采用收益、置信度、简洁性等来衡量所发现知识的有效性、可用性和可理解性。

⑤使用场合不同。Web 信息检索用于用户有明确请求的场合，当用户没有明确的信息需求，或信息检索系统返回太多的结果以致用户无法一一浏览时就需要使用 Web 挖掘技术。

尽管 Web 挖掘是比 Web 信息检索层次更高的技术，但它并不是用来取代信息检索技术的，二者是相辅相成的。一方面，这两种技术各有所长，有各自适用的场合；另一方面，我们可以利用 Web 挖掘技术来提高信息检索的精度和效率，改善检索结果的组织，使信息检索技术发展到一个新的水平。

（2）Web 数据挖掘与信息抽取

【定义 1-3】

信息抽取指从给定的文档中抽取特定类别的信息，并以结构化形式对这些信息进行组织和表示的技术。

文档可以是结构化、半结构化及非结构化。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针对特定网站，一些抽取方法能够

① 刘振岩，王万森，陈立平. Web 信息检索与 Web 数据挖掘［J］. 微机发展，2003（7）：66-68.

自动或半自动地建立抽取模式（Kushmerich，1999），对于这类信息抽取，Web 数据挖掘可以看做是一个信息抽取的过程。此外，在 Web 数据挖掘中，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 数据挖掘的预处理过程，即在对 Web 上的文本信息进行挖掘之前，通过信息抽取技术可以将其中的非结构化数据转化为结构化数据，然后再对其实施传统的挖掘算法进行知识发现。信息抽取技术并不试图全面理解整篇文档，如文本意义的细微差别以及作者的写作意图等深层理解问题，而只是对文档中用户感兴趣的事实信息部分进行收集和分析。

1.1.3 Web 数据挖掘的流程

传统数据挖掘是 Web 数据挖掘的基础，因此，传统数据挖掘与 Web 数据挖掘在流程上有相通之处，但是，由于 Web 挖掘本身的特点，决定了具体的挖掘过程又有所区别。典型的 Web 数据挖掘包括四个步骤，如图 1-1 所示。

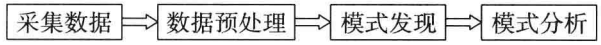


图 1-1 Web 数据挖掘的流程

1. 采集数据

采集数据即从外部的 Web 环境中有选择地获取数据，为后面的数据挖掘提供资源。通常，采集数据由数据搜索、数据选择和数据收集等 3 个独立的过程组成。

Web 数据挖掘的主要数据源如下：

（1）服务器日志

包括访问日志和引用日志。访问日志记录 Web 浏览中点击以及每次执行成功或失败的请求。引用日志是 Web 服务器上的日志文件，包含访问者的访问位置和引入 Web 站点的关键词或路径。

（2）Cookie

Cookie 是用户访问站点时由 Web 服务器传递到用户浏览器的少量信息，详细描述了访问者访问站点时浏览了哪些地方，当访问者下次再访问同一网站时，Cookie 会自动识别用户。利用 Cookie 可以跟踪统计用户访问该网站的习惯，比如什么时间访问，访问了哪些页面，在每个网页的停留时间等。利用这些信息，一方面可以确定访问者的身份和偏好，为用户提供个性化的服务；另一方面，也可以作为了解所有用户行为的工具，对于网站经营策略的改进有一定参考价值。

（3）表单或用户注册数据

主要是访问者在进入站点时注册提供的个人信息，如姓名、地址、出生日期、性别以及职业等，为 Web 挖掘提供重要的数据。

（4）电子商务站点交易数据

电子商务站点的交易数据记录了大量的客户历史交易数据，根据历史交易记录，可以挖

掘客户的行为模式和兴趣爱好，从而向客户推荐相关的个性化商品，提高客户的满意度。

2. 数据预处理

主要对数据采集所获得的源数据进行加工处理和组织重构，即从源数据集中除去明显错误和冗余的数据，进一步精简所选数据的有效部分，将数据转换成有效形式，同时构建相关主题的数据仓库，为下一步的数据挖掘过程创建基础平台。与传统数据挖掘一样，数据预处理是为数据挖掘所做的前期准备，它主要包括数据清理、数据集成、数据转换和数据约简等，但由于 Web 上数据的特点，具体操作过程会有所不同。

（1）数据清理

在这一阶段中，将要完成的任务主要是去除源数据中的噪声和无关数据，处理遗漏数据和清洗脏数据，包括重复数据处理和缺值数据处理等，并且完成一些数据类型的转换。例如 Web 数据挖掘一般要去除 ROBOT 或 SIPDER 请求以及一些错误请求等。

（2）数据集成

数据集成是将上阶段清洗过的数据进行集成、归类。例如，在研究用户浏览模式的日志记录中，需要识别每一位用户的浏览记录以及每一位用户的不同会话时段。因此，必须对采集的数据记录根据用户和会话时段的不同进行数据的归类集成。

（3）数据转换与数据约简

数据转换就是将数据转换成适合进行数据挖掘的形式。数据约简是在对挖掘任务和数据内容充分理解的基础上，通过寻找数据的有用特征，在尽可能保持数据信息原貌的前提下，最大限度地精减数据量，提高数据挖掘的算法效率。

3. 模式发现

模式发现是数据挖掘系统的核心部分，主要是运用各种数据挖掘技术，从海量数据中提取出潜在的、有效的且能被人理解的知识模式，而这海量数据是经过以上预处理后的数据。Web 数据挖掘结合传统数据挖掘技术和 Web 挖掘技术来进行模式发现。

4. 模式分析

对发现的模式进行解释和评估，必要时需返回前面处理中的某些步骤以反复提取，最后，将发现的知识以能理解的方式提供给用户，可以是机器自动完成，也可以是与分析人员进行交互来完成。

1.1.4　Web 数据挖掘的分类[①]

Web 信息的复杂性决定了 Web 数据挖掘任务的多样性。根据挖掘对象不同，可以将 Web 数据挖掘分为三类（如图 1-2 所示）：Web 内容挖掘、Web 结构挖掘和 Web 使用挖掘。

1. Web 内容挖掘

Web 内容挖掘是一种基于网页内容的 Web 挖掘，是从大量的 Web 数据中获取潜在的、

① 刘树超，等. Web 数据挖掘研究与探讨［J］. 制造业自动化，2010（9）：163-166.

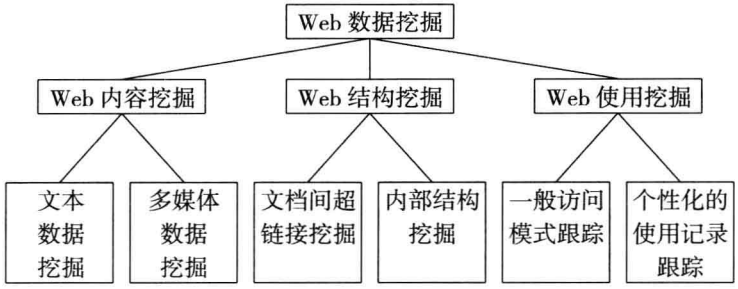


图 1-2 Web 数据挖掘的分类

有价值的知识或模式的过程，是对网页上真正的数据进行挖掘，包括网页内容挖掘和搜索结果挖掘。从网络信息源形式看，大量网络信息资源可以直接从网上抓取、建立索引、实现检索服务。从资源形式来看，网络数据对象，既有文本和超文本数据，也有图形、图像、语音等多媒体数据；既有来自于数据库的结构化数据，也有用 HTML 标记的半结构化数据和无结构的自由文本。

Web 内容挖掘针对的对象是文本文档和多媒体文档，就其挖掘内容而言，Web 内容挖掘又可以分为 Web 文本数据挖掘和 Web 多媒体数据挖掘。Web 文本数据挖掘，是 Web 内容挖掘中比较重要的技术领域，可以对 Web 上大量的文档集合的内容进行总结、分类、聚类和关联分析等。Web 多媒体数据挖掘包括运用挖掘技术对 Web 上的音频、视频和图像数据进行挖掘，目前还处于探索阶段。

（1）Web 文本数据挖掘

以计算语言学、统计树立分析为理论基础的 Web 文本挖掘，是从大量的文本数据中发现和提取隐含的、事先未知的知识，最终形成用户可以理解的、有价值的信息和知识过程。

按照文本数据挖掘的对象又可以把文本挖掘分为：基于单文档的数据挖掘和基于文档集的数据挖掘。文本挖掘的结果可以是对某个文本内容的总结概括，也可以是对整个文本集合的分类或聚类结果。从方法上看，许多传统的挖掘算法经过修改后都可以用在 Web 文本挖掘上。一般来说，基于文本的 Web 挖掘方法有数据库方法、建立 Web 数据仓库方法和新近的基于软件 Agent 的分类器方法、基于概念的文本信息挖掘方法等。从功能上看，Web 文本挖掘主要是应用这些算法对文档集合进行分析以及利用 Web 文档进行趋势分析等。

（2）Web 多媒体数据挖掘

Web 多媒体数据挖掘就是从大量多媒体集中，通过综合分析视听特性和语义，发现隐含的、有效的、有价值的、可理解的模式，进而发现知识，得出事件的趋向和关联，为用户提供问题求解层次的决策支持能力。

目前，Web 多媒体数据挖掘主要研究的是多媒体文本内容的挖掘，也研究其结构，因

此，多媒体数据挖掘又分为多媒体文本内容挖掘和多媒体文本结构挖掘。多媒体文本内容挖掘应用于[①]：①提取多媒体文本文档中的中心词汇，并以此为主对多媒体文档进行文本总结；②根据多媒体文本上下文内容进行翻译。多媒体文本结构挖掘可应用于文件格式的挖掘和研究。

由于多媒体数据挖掘的对象是图像、视频和音频等多媒体数据，所以，如时间、空间和视听等内容特征，和一般关系数据库中数据的特征在许多方面都不同，因此，一些传统的数据挖掘方法不能直接采用，需要研究适合于多媒体数据的新的挖掘方法和技术。目前，从方法上看，基于多媒体信息挖掘通常采用关联规则法和特征提取法。从功能上看，Web多媒体挖掘主要是通过对Web上的音频、视频数据和图像进行预处理，运用挖掘技术挖掘其中潜在的、有价值的信息和模式。近年来，基于多媒体的信息挖掘引起了许多研究人员的关注，但是目前，多媒体数据的挖掘研究还在处于探索阶段。

2. Web结构挖掘

Web结构挖掘是从Web结构和链接关系中推导出潜在知识和模式的过程。其中，Web结构包括不同网页之间的超链接结构和一个页面内部的树形结构，以及文档URL中的目录路径结构等。通过对这些站点结构进行分解、变形和归纳，可以将页面进行分类和聚类，并且帮助找到权威页面以及中心页面，从而提高信息检索效率。链接关系是指不同Web文本之间的链接关系，它反映了文档之间的引用关系，而这些关系数据背后蕴含着丰富有价值的信息，利用这些可以对所找到的Web页面进行排序，找出最重要的页面。

Web结构挖掘是针对链接信息这一重要的Web数据，它可以从Web的组织结构以及引用和被引用间的链接关系中发现许多蕴含在Web页面之外的对我们有潜在价值的模式和知识，利用这些知识可以对页面进行排序，发现重要的、权威的页面等。

Web结构挖掘和Web内容挖掘都是对Web上第一类数据，即真正的原始数据进行挖掘，因此这两种挖掘任务经常一起使用。同时，Web结构挖掘在网站优化中有着重要作用，可以评价和分析网页的质量，也有助于优化网页的链接设计，减少不合理的链接，从而改进网站结构。同时还可以用来指导网页采集工作，提高采集效率等。

Web结构挖掘的基本思想是将Web看成一个巨大的以页面为节点、页面之间超链接为有向边所构成的一个网状结构的有向图，然后利用图论对Web的拓扑结构进行分析，从而可以发现重要页面和权威页面，以确定网站结构的合理性。

Web结构挖掘的算法一般可分为查询无关算法和查询相关算法两类。PageRank算法和HITS（Hypertext Induced Topic Search）算法分别是查询无关算法和查询相关算法的代表。这些算法已经在实际的系统中实现和使用，并且取得了良好的效果。在接下来的一章中我们会重点介绍这两种算法。

① 涂承胜，鲁明羽，陆玉昌. Web内容挖掘技术研究［J］. 计算机应用研究，2003（11）：5-9.