国家社科基金
GUOJIA SHEKE JIJIN HOUQI ZIZHU XIANGMU
后期资助项目

数据挖掘在Web资源开发与利用中的应用研究

Application Research of Data Mining in the Development and Utiliazation of Web Resources

刘忠宝 著

科学出版社

国家社科基金
GUOJIA SHEKE JIJIN HOUQI ZIZHU XIANGMU
后期资助项目

数据挖掘在Web资源开发与利用中的应用研究

Application Research of Data Mining in the Development and Utiliazation of Web Resources

刘忠宝 著

科学出版社
北京

内 容 简 介

本书为国家社科基金后期资助项目成果，针对 Web 资源开发利用面临的主要问题，围绕数据挖掘优化方法，对用户行为分析、个性化推荐、Web 信息检索以及 Web 页面链接分析等方面的内容展开研究。该成果定性与定量研究、理论与实证研究相结合，融合多个学科的技术成果，在研究方法和手段上有所创新。该成果既有翔实的理论阐述，又有系列的公式演示，严谨可信，具有较高的理论研究价值；同时该成果提出的一些新型模型和理论框架具有较高的应用价值。

本书对信息资源领域专家具有一定的比较借鉴价值，适合作为图书馆学、情报学、计算机科学等学科科研人员和研究生的参考用书。

图书在版编目(CIP)数据

数据挖掘在 Web 资源开发与利用中的应用研究 / 刘忠宝著. —北京：科学出版社，2016.12

ISBN 978-7-03-051401-1

Ⅰ.①数… Ⅱ.①刘… Ⅲ.①数据采集-应用-互联网络-资源开发-研究②数据采集-应用-互联网络-资源利用-研究 Ⅳ.①TP393.4

中国版本图书馆 CIP 数据核字（2016）第 313726 号

责任编辑：刘 超 / 责任校对：张凤琴

责任印制：张 伟 / 封面设计：无极书装

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

北京教图印刷有限公司 印刷

科学出版社发行 各地新华书店经销

*

2016 年 12 月第 一 版 开本：720×1000

2016 年 12 月第一次印刷 印张：14

字数：244 000

定价：88.00 元

（如有印装质量问题，我社负责调换）

国家社科基金后期资助项目
出版说明

后期资助项目是国家社科基金项目主要类别之一，旨在鼓励广大人文社会科学工作者潜心治学，扎实研究，多出优秀成果，进一步发挥国家社科基金在繁荣发展哲学社会科学中的示范引导作用。后期资助项目主要资助已基本完成且尚未出版的人文社会科学基础研究的优秀学术成果，以资助学术专著为主，也资助少量学术价值较高的资料汇编和学术含量较高的工具书。为扩大后期资助项目的学术影响，促进成果转化，全国哲学社会科学规划办公室按照“统一设计、统一标识、统一版式、形成系列”的总体要求，组织出版国家社科基金后期资助项目成果。

全国哲学社会科学规划办公室

2014 年 7 月

目　录

第一章 绪 论

第一节 数据挖掘研究进展

随着互联网技术的不断发展，网上的数据量日益增长，人们往往在数据海洋中迷失方向。如何对海量数据进行分析并发现有用知识成为时下人们关注的热点问题。数据挖掘技术的出现为人们解决上述困扰提供了可能。数据挖掘是指通过对海量的数据进行有目的地提取、分拣、归类，挖掘出有用信息，用于为行业领域提供决策支持。当前主流的数据挖掘技术主要包括特征降维、智能分类、聚类分析等三方面的内容。

一、特征降维

真实世界中的很多数据是高维的，即数据包含很多属性或特征。尽管高维数据比低维数据拥有更多的信息量，但在实际应用中对高维数据进行直接操作将会非常困难。首先，“维数灾难”[1,2]会导致分类学习所需的有标记样本随着维数的增加，计算量呈指数倍增长，部分算法在极高维空间中甚至无法工作；其次，人们在低维空间中形成的一些直觉在高维空间中可能会失效。例如，对于二维空间中的单位圆和单位正方形，两者的面积相差不多；对于三维空间中的单位球和单位立方体，两者的体积也差不多。然而随着维数的升高，在高维空间中超球的体积相对于超立方体的体积会迅速变小。

为了解决高维数据所面临的问题，一种有效的方法是对其进行降维。笼统地说，降维是指为高维数据获取一个能忠实反映原始数据固有特性的低维表示。降维有特征选择和特征提取两种方式[3-5]。

特征选择的基本原则是选择类别相关的特征而排除冗余的特征。即根据某种准则从一组数量为 D 的特征中选择出数量为 d（$D>d$）的一组最优特征的过程。特征选择通过降低原始数据的相关性和冗余性，在一定程度上解决了“维数灾难”问题。特征选择主要分三类[6-9]：①过滤法。设计一个评分函数对每个特征评分，按分数高低将特征排序，并将

排名靠前的特征作为特征子集；②封装法。把学习机作为一个黑箱并通过验证样本上的正确率来衡量特征子集的性能，一般采用向前或向后搜索生成候选特征子集；③嵌入式法。该方法是一种结合学习机评价特征子集的特征选择模型，具有封装法的精度和过滤法的效率。近年来，众多学者从事特征选择研究，并取得一些成果。Kira 和 Rendell 提出的 Relief[10]算法根据特征值在同类实例中以及相近不同类实例中的区分能力来评价特征的相关度；Nakariyakui 和 Casasent 提出的分支跳跃法[11]通过对解决方案树中某些节点不必要评价函数的计算来提高搜索速度；Brodley 提出的 FSSEM 算法[12]根据最大化的期望值来选择特征子集；Whiteson 等提出的 FS-NEAT 算法[13]通过特征集合搜索和拓扑网络学习解决特征选择问题。

特征提取是指原始特征空间根据某种准则变换得到低维投影空间的过程。与特征选择相比，特征提取的降维效率更高[14]。特征提取可分为线性方法和非线性方法两类。经过几十年的发展，研究人员提出多种线性特征提取方法：非负矩阵分解（non- negative matrix factorization，NMF）[15]通过将原始特征空间低秩近似保证降维后的特征非负；因子分解（factor analysis）[16]通过降低原始特征空间的相关性实现降维；奇异值分解（singular value decomposition，SVD）[17]通过考察奇异值的贡献率实现降维；主成分分析（principal component analysis，PCA）[18]通过对原始特征空间方差的研究得到一组正交的主成分；独立成分分析（independent component analysis，ICA）[19]利用原始特征空间的二阶和高阶统计信息，进一步提高了 PCA 的降维效率；线性判别分析（linear discriminant analysis，LDA）[20]通过最大化类间离散度和类内离散度的广义 Rayleigh 熵实现特征变换。线性特征提取方法不能保持原始特征空间的局部信息，没有充分考虑数据的流形结构。鉴于此，近年来出现了众多非线性特征提取方法：核主成分分析（kernel principal component analysis，KPCA）[21]和核 Fisher 线性判别分析（kernel linear Fisher discriminant analysis，KLDA）[22,23]分别在 PCA 和 LDA 的基础上引入核方法，将 PCA 和 LDA 的适用范围从线性空间推广到非线性空间；多维缩放（multi-dimensional scaling，MDS）[24]通过保持数据点间的相似性实现降维；ISOMAP（isometric mapping）的主要思想是利用数据间的测地线距离代替欧式距离，然后利用 MDS 来求解；局部线性嵌入（locally linear embedding，LLE）[25]利用稀疏矩阵算法实现降维；拉普拉斯特征映射

(Laplacian eigenmap, LE)[26]利用谱技术实现降维。此外，近年来还出现了一系列基于流行学习的算法，如局部切空间排列方法（local tangent space alighment, LTSA)[27]、海森特征谱方法（Hessian eigenmap, HE)[28]、保局投影（locally preserving projections, LPP)[29]、近邻保持映射（neighborhood preserving projection, NPP)[30]等。这些算法本质上都是非线性降维方法，并没有利用样本的类别信息，鉴于此，研究人员提出了有监督非线性降维方法，如判别近邻嵌入（discriminant neighborhood embedding, DNE)[31]、最大边缘投影（maximum margin projection, MMP)[32]等。DNE 算法基于不同类别拥有不同低维流形这一假设，为每个类别分别建立流形结构，然后通过最大化不同类别间近邻样本距离、最小化相同类别近邻样本距离得到最终的子空间。MMP 则是一种半监督学习方法，其考虑是现实中获得样本的标记比较困难，所以对于获得的样本，如果有标记则应尽量区分不同的流形结构，如果没有标记则尽量发现其所在的流形结构。MMP 将 LPP 和 DNE 结合，通过求解广义特征值问题得到子空间。流形学习在数据可视化领域得到了广泛的应用。然而，由于流形学习隐式地将数据从高维空间向低维空间映射，所以其不足之处在于无法得到新样本在低维子空间的分布。流形学习在描述邻域结构时还存在邻域选择、邻域权重设置等问题。

上述降维方法无法解释各变量对数据表示和分类的影响。鉴于此，研究人员提出基于稀疏表示的特征提取方法。稀疏表示是由傅里叶变换和小波变换等传统的信号表示扩展而来，目前在模式识别、计算机视觉、信号处理等领域得到成功的应用。迄今为止，基于稀疏的降维方法典型代表有[33]：稀疏主成分分析（sparse principal component analysis, SPCA)、稀疏线性判别分析（sparse linear discriminant analysis, SLDA)、稀疏表示分类器（sparse representation-based classification, SRC)、稀疏保持投影（sparsity preserving projections, SPP）等。SPCA 没有考虑样本的类别信息，因此，不利于后续的分类任务。SLDA 可以用于解决二分类问题，但对于多类问题，并不能像 LDA 那样进行直接的扩展。SRC 在流形稀疏表示的框架下保持数据的局部属性，该方法成功地应用于人脸识别。SPP 将系数表示的稀疏性作为一种自然鉴别信息引入到特征提取中，并在人脸数据集上证明了其有效性。然而，稀疏表示在寻找子空间的过程会牺牲类内的同一性，因为它是对单个样本分别获取它们的稀疏表示，因此缺乏对数据全局性的约束，无法准确地描述数据的全局结构。当数据包含大量噪声

或者有损坏时，这个缺点会使算法的性能明显下降。

二、智能分类

智能分类是数据挖掘的另一项重要内容，分类技术的核心是构造分类器。分类器一般具有良好的泛化能力，能够准确地预测未知样本的类别。分类器工作一般经历训练和测试两个阶段。训练阶段根据训练数据集的特点得到分类标准；测试阶段完成新进数据类属判定的任务。按照不同的标准，可对分类器进行如下分类：根据工作原理，可将分类器分为概率密度模型、决策边界学习模型和和混合模型。概率密度模型在估计每类概率密度函数的基础上用贝叶斯决策规则实现分类；决策边界学习模型在学习过程中最优化一个目标函数，该函数表示训练样本集上的分类错误率、错误率的上界或与分类错误率相关的损失；混合模型先对每类模型建立一个概率密度模型，然后用判别学习准则对概率密度模型的参数进行优化。根据表达形式，可将分类器分为区分模型和生成模型。区分模型通过对训练样本学习生成分类标准；生成模型根据概率依赖关系构造分类模型。根据求解策略，可将分类器分为基于经验风险最小化模型和基于结构风险最小化模型。早期的分类器求解算法基本上基于经验风险最小化原则；结构风险最小化模型基于权衡经验风险和置信范围。

近年来，智能分类受到中外学者的极大关注，在数据挖掘、机器学习、情报分析等领域得到广泛研究并取得令人振奋的成果。决策树分类方面，Quinlan 提出的 ID3 算法[34]在信息论互信息的基础上建立树状分类模型；针对 ID3 的不足，先后提出 C4.5[35]、PUBLIC[36]、SLIQ[37]、RainForest[38]等改进算法。基于关联规则分类方面，Liu 等提出的关联分析算法（classification based on association，CBA）[39]采用经典的 Apriori 算法发现关联规则；Li 等提出的多维关联规则的分类算法（classification based on multiple class association rules，CMAR）[40]利用 FP-Growth 算法挖掘关联规则；Yin 等提出的预测性关联规则分类算法（classification based on prediction association rules，CPAR）[41]采用贪婪算法直接从训练样本中挖掘关联规则。支持向量机方面，Vapnik 等提出支持向量机（support vector machine，SVM），由于最优化问题中有一个惩罚参数 C 因此也称为 C-SVM[42-44]；由于参数 C 没有确切含义且选取困难，Scholkopf 等提出 ν-SVM[45]，其中参数 ν 用来控制支持向量的数目和误差且易于选取；通过扩展 SVM 最大间隔的思想，Scholkopf 在前人工作的基础上提出单类支持

向量机（one class support vector machine，OCSVM）[46]，该方法通过构造超平面来划分正常数据和异常数据；针对单类问题，Tax 等提出支持向量数据描述（support vector data description，SVDD）[47]的概念，该方法采用最小体积超球约束目标数据达到剔除奇异点的目的；Tsang 等提出基于最小包含球（minimum enclosing ball，MEB）的核心向量机（core vector machine，CVM）[48]，该方法有效地提高了 SVM 求解二次规划问题的效率。此外，常见的 SVM 变种还有：最小二乘支持向量机（least squares support vector machine，LSSVM）[49]、Lagrangian 支持向量机（largrangian support vector machine，LSVM）[50]、简约支持向量机（reduced support vector machine，RSVM）[51]、光滑支持向量机（smooth support vector machine，SSVM）[52]等。贝叶斯分类方面，Kononenko 提出的半朴素贝叶斯分类器（semi-naive Bayesian classifier）[53]采用穷尽搜索的属性分组技术实现分类；Langley 等提出的基于属性删除的选择性贝叶斯分类器（selective Bayesian classifier based on attribute deletion）[54]通过删除冗余属性来提高分类精度；Kohavi 通过将朴素贝叶斯分类器和决策树相结合提出朴素贝叶斯树型学习机（naïve Bayesian tree learner，NBT）[55]；Zheng 等提出的基于懒惰式贝叶斯规则的学习算法（lazy Bayesian rule learning algorithm，LBR）[56]将懒惰式技术应用到局部朴素贝叶斯规则的归纳中；Friedman 等提出的树扩张型贝叶斯分类器（tree augmented Bayesian classifier，TAN）[57]通过构造最大权生成树实现分类。此外，还有神经网络分类算法、K-近邻分类法、基于粒度和群的分类算法等。

上述分类方法各有特点和适用范围，它们之间互相渗透、相互共存。经过几十年的发展，智能分类方法表现出强大的生命力，其理论体系不断完善，应用领域不断扩大，关注程度不断提高。随着相关理论和技术逐步完善，智能分类理论和方法必将不断发展。

三、聚类分析

聚类分析是指将一个数据集依据某种规则分成若干子集的过程，这些子集由相似元素构成。聚类分析是一种典型的无监督学习方法，它在进行分类与预测时无需事先学习数据集的特征，具有更优的智能性。聚类分析在 Web 资源开发与利用中发挥着重要作用。

当前主流的聚类算法包括以下几类：层次聚类算法、划分聚类算法、基于密度和网格的聚类算法以及其他聚类算法。

层次聚类算法利用数据的连接规则，通过层次架构的方式反复将数据分裂或聚合，以便形成一个层次序列的聚类问题的解。典型代表有：Gelbard 等提出的正二进制方法[58]，该方法将待聚类数据存储在以由 0 和 1 组成的二维矩阵中，其中行表示记录，列表示属性值，1 和 0 分别表示记录是否存在对应的属性值；Kumar 等提出基于不可分辨粗聚合的层次聚类算法（rough clustering of sequential data，RCOSD）[59]，该算法适用于挖掘连续数据的特征，可以帮助人们有效地描述潜在 Web 用户组的特征；此外，基于 Quartet 树的快速聚类算法[60]以及 Hungarian 聚类算法[61]也具有一定代表性。层次聚类算法最大优势在于无需事先给定聚类数量，可以灵活地控制聚类粒度，准确表达聚类簇间关系。主要不足在于其无法回溯处理已形成的聚类簇。

划分聚类算法需要事先给出聚类数量或聚类中心，为了确保目标函数最优，不断迭代，直至当目标函数值收敛时，可得聚类结果。典型代表有：MacQueen 提出的 K-means 算法[62]，该算法试图找到若干个聚类中心，通过最小化每个数据点与其聚类中心之间的距离之和来构建最优化问题；为了提高 K-means 算法的普适性，Huang 提出了面向分类属性数据的 K-modes 聚类算法[63]；Chaturvedi 等提出面向分类属性数据的非参数聚类方法 K-modes-CGC[64]；Sun 等在 K-modes 算法基础上提出迭代初始点集求精 K-modes 算法[65]；Ding 等将统计模式识别中的重要概念——最近邻一致性应用到聚类分析提出一致性保留 K-means 算法[66]；Ruspini 将模糊集理论与聚类分析有机结合起来，提出模糊聚类算法（fuzzy c-means，FCM）。划分聚类算法的优点在于收敛速度快，缺点是该类算法需要事先指定聚类数量。

基于密度的聚类算法利用数据密度发现类簇；基于网格的聚类算法通过构造一个网格结构实现模式聚类。上述两类算法适用于空间信息处理，并常常合并在一起使用。典型代表有：Zhao 等提出的网格密度等值线聚类算法（grid-based density isoline clustering，GDILC）[67]；Ma 提出的基于移位网格的非参数型聚类算法（shifting grid clustering，SGC）[68]；Pileva 等提出的面向高维数据的网格聚类算法[69]；Micro 等提出基于密度的自适应聚类算法[70]，该算法适用于移动对象轨迹数据处理，并且对处理形状复杂的簇具有明显的优势。其他一些常见的聚类算法有：Tsai 等提出一种新颖的具有不同偏好的蚁群系统[71]，该系统用以解决数据聚类问题；基于最大 θ 距离子树的聚类算法、图论松弛聚类（graph-based

relaxed clustering，GBR）算法以及基于 dominant 集的点对聚类算法。

随着互联网技术的发展，Web 资源规模日益增大，各种结构复杂的数据不断涌现。如何对这些复杂数据进行聚类分析成为广大研究人员面临的重要课题。

第二节 Web 资源开发与利用进展

Web 资源是一种新型的数字化资源和社会化信息。它利用超文本链接，构成立体网状文献信息链，把不同国家、不同地区、各种服务器、各种网站网页、各种不同信息资源通过节点连接起来，形成多渠道、互动交流的“非出版的数字化信息”。与传统信息资源相比，Web 资源无论在数量类型、分布结构、传播范围、交流机制、查询手段等方面都存在明显差异。Web 资源开发与利用主要包括 Web 资源保存、组织管理和信息提取三方面的内容。

一、Web 资源保存

Web 资源数量大、种类多、生命周期短、更新速度快，若不对其采取保存措施，将会造成重要数字资源的丢失。国内外目前已经达成了共识，即 Web 资源是国家文化的重要组成部分，必须对其进行长期保存[72,73]。近年来，世界各国开展了一些重要的 Web 资源保存项目，典型的代表如下。

（一）美国国会图书馆 Minerva 项目

美国国会图书馆启动 Minerva 项目旨在收集和保存原生网络信息资源。美国国会拨款 1 亿美元给国会图书馆进行数字保存项目研究，并委托国会图书馆制定国家数字信息基础结构和保存项目的具体计划。同时敦促国会图书馆和其他联邦机构、研究界、图书馆界及商界等部门密切合作。该项目主要对 6 个方面的内容进行保存，即 Web 信息、数字视频、数字音频、数字期刊、电子图书和数字电视。其中，Web 资源的整理和保存被视为该计划的重要组成部分。

（二）澳大利亚国家图书馆的 Pandora 项目

澳大利亚国家图书馆启动了 Pandora 计划，与其他 9 家单位合作形成

了分布式的保存网络，以确保所选择的澳大利亚网络出版物及其他信息可长期读取。这些 Web 资源是澳大利亚文化遗产的重要组成部分。该项目采取有选择地保存网络信息资源的策略，并把所采集的 Web 资源按内容分为 15 大种类：艺术和人文、商业与经济、计算机与网络、教育、环境、健康、历史与地理、本地居民、青少年读物、法律与犯罪、新闻与媒体、政治与政府、科技、社会与文化、运动与娱乐等。作为国际互联网保存协会的成员，澳大利亚国家图书馆积极推动 Pandora 项目参与国际合作，以探讨和共享与 Web 资源归档保存有关的技术。

（三）英国 Web 资源保存项目

英国国家图书馆的 Britain On The Web 项目有选择地收集了 Web 资源 9 大种类：国防与对外政策、司法与国家安全、环境、就业和国家财政、健康和教育、文化媒体和体育、市场服务政策、政府与宪法筹备、公众调查，同时提供免费的服务。此外，英国 Web 存档联盟项目的目标是研究存档解决方案，以保证在网络空间内出版的有价值的学术、文化和科学资源不会丢失。包括大英图书馆在内的 6 个机构参与了该项目。

（四）欧洲的 NEDLIB 计划

NEDLIB 是欧洲国家图书馆的合作项目，该项目由荷兰国家图书馆领导，参加方包括法国、挪威、芬兰、德国、葡萄牙、瑞士、意大利等国的国家图书馆，以及 Kluwer Academic、Elsevier Science、Springer Verlag 等 3 家出版机构。其目标是建立欧洲网络化存储体系的基础构架，并致力保证所收集与保存的电子出版物可在现代和未来使用。

（五）瑞典国家图书馆 Kulturarw 项目

Kulturarw 项目的目的是测试瑞典在线文献的收集、保存和提供读取的方法。至今已经成功地完全下载 7 个瑞典网站，收藏约 6500 万条信息，数据量达到 300GB，其中有一半是文本文件，主要是 HTML 和纯文本格式。

（六）法国 BnF 网络保存计划

该计划由法国国家图书馆负责实施，旨在存储和管理网络文献，为未来提供特定历史时期具有代表性的 Web 资源。BnF 采取选择性保存策

略，对于正规的网络出版物，采取人工选择，但该方法效率较低。而对于更为广泛的 Web 资源，则用自动爬行器来获取。自动爬行程序的使用使管理更广范围的网址成为可能，而且可以最快最大量地收集到网上巨大规模的文献信息。

（七）德国海德堡大学的汉学研究数字档案馆项目

该项目是汉学研究数字信息资源欧洲中心的一部分，位于德国海德堡大学汉学研究所内。除了提高本地所有印刷资源的质量以及改善获取渠道外，其宗旨是进一步推动欧洲各地获取和利用汉学相关数字信息资源。为此，该中心收集各种形式的全文数据库，并使尽可能多的人可以存取；编制虚拟图书查询检索系统和联合目录，帮助找出欧洲各图书馆所收藏的与中国有关的印刷资源；创办重要的有关中国的网络资源指南；建设符合上述宗旨的 IT 基础设施。

（八）中国国家图书馆的 WICP 项目

中国国家图书馆组织实施了网络信息资源保存的实验项目。该项目采用了两种保存策略：一是镜像存档，即以网站为信息单元进行网络信息存档。从对象网站的首页开始，收集该网站的全部网页信息，采集的数据保持原来的目录结构，并保存到存档系统中。在不同的时间点对同一对象做重复采集，即形成该对象网站的时间切片。对网站进行编目，元数据输入到国家总书目中；二是专题存档，即以网页为信息单元进行网络信息存档。按不同专题确定对象网站，从对象网站的首页开始，收集该网站下的有用网页，进行内容提取、自动分类和标引，并将其保存到存档系统。

（九）中国 Web 信息博物馆

中国 Web 信息博物馆由北京大学计算机网络与分布式系统实验室主持开发的中国网页历史信息存储与展示系统，包括历史网页存储系统和回放系统两个部分。这两部分独立完成各自的任务，回放系统是基于存储系统完成的。目前系统可以收集中国所有静态网页，并提供历史网页的存档和回放。该系统主要有如下几项功能：① 输入 URL，浏览永久保存的历史网页；② 典型历史网页展示，可以顺着超链接在历史网页中持续浏览；③ 历史事件专题回放。与普通网上搜索不同的是，它能为用户

提供一个完整的历史网页，而不是单篇文章。这对于追寻重大历史事件发展进程的全貌有着特殊意义。作为全国最大、最完整的互联网内容信息收集与仓储中心，中国 Web 信息博物馆现收藏有约 10 亿个中文网页，并以平均每月增加 1000 万网页的速度扩张。

二、Web 资源组织

Web 资源的组织建立在超文本传输协议 HTTP 和超文本语言 HTML 基础之上，以超文本或超媒体的形式实现。Web 上的信息通过 Web 站点上的页面表示，简称为 Web 网页。这些页面使用 HTML 语言，并利用 HTML 标记和各种多媒体资源构成链接，形成超文本，同时还可以通过标记链接到本站点或其他站点的页面，形成数据网络。Web 网页包括主页和子页，是互联网的基本信息组织方式，也是用户使用网络信息资源的主要形式。

当前 Web 资源组织方法主要有元数据法、内容分类法、主题组织法等方法[74]。

（一）元数据法

元数据法包括数据库和搜索引擎。元数据是对资源的信息进行描述，即关于数据的数据，目前常见的类型有 MARC、GILS、TEI、FGDC、DC、IAFA 等。从元数据的定义中可以看出，元数据法可以帮助用户更好地识别、评价、引用 Web 资源。在对 Web 资源的组织上，元数据法所起到的作用有：知识描述（描述 Web 资源的内容、主题、关键词等，利于用户了解 Web 资源的中心内容）、知识定位（提供 Web 资源的来源、存储位置等）、语义搜索（提供链接或其他便于找到 Web 资源信息）、知识评估（对 Web 资源的价值、准确性、权威性等进行评估）等。应用元数据法，可以对 Web 资源的现有一般信息即内容和主题等进行更深入的描述，如其出版信息、作者信息、合作信息、时效性、效益性等，通过以上信息的描述和整理，用户就能更便捷地对 Web 资源进行组织，判断其能否适合自己的解决问题的需要，能为自己创造什么利益，进而做出选择。

然而，由于元数据法尚处于起步阶段，没能形成完善的组织系统，在很多方面都存在这样那样的问题，所以急需研究者对之进行细化和完善，使之发挥更大的作用。

（二）内容分类法

内容分类法是指对从网络上得到的各学科领域的信息资源进行初步的识别后，进一步整理、归纳，按照详细内容进行分类，把相同领域的信息整合到一起，并按一定的顺序和规律进行系统的索引编号。内容分类法是传统图书馆对馆藏书籍进行分类、收编的方法之一，如今随着网络信息技术的发展，大量的网络信息和电子书籍不断涌现，这一传统方法也逐渐被应用于 Web 资源，并借助于现代信息技术中数据库、搜索引擎的帮助，焕发出新的生机。内容分类法的优势在于其将繁杂的庞大信息资源分门别类，使原本错综复杂的信息资源变得清晰条理，犹如为用户编绘了一张简单明了的索引图，用户只要按照索引提示“按图索骥”，就能方便地得到自己想要的信息资源，也能方便地对其进行评价。

内容分类法将所有信息资源按某种事先确定的体系结构组织信息，用户通过逐层浏览选择信息，具有严密的系统性和良好的可扩充性，但不适用于建立大型综合性资源系统，仅在建立专业性或示范性信息资源体系时才显出结构清晰和使用方便的优点。在网络环境下组织信息资源必须对内容分类法进行改造，要增补类目，增强语言的直观性和透明度，扩展同主题法即主题词表的联系，增强兼容性和国际通用性。

（三）主题组织法

主题组织法按照确定的规范标准对现有信息按其主题（专业、领域等）建立起一个大的主题，然后在这大主题下，继续细分，建立起一系列平行的小主题目录，最后形成树状的主题树网络。这样做的好处是：① 揭示信息直观，事实上，主题词也是组织 Web 资源的重要方法之一；② 便于特性检索。主题组织法从信息本身特质出发，结合搜索手段，能最便捷地根据信息的特性检索到目标信息；③ 设置浏览等级。主题组织法可以根据不同用户的知识水平和需求级别，设置不同的信息等级，用户可以根据自身实际情况，直接跳过不适合的等级进入相应等级的信息库中检索信息，从而节省了精力和时间。主题组织法建立起后，以其严谨、指向性强、检索准确性高、系统严密等特点迅速成为 Web 资源组织的热点。从目前来看，主题组织法的作用体现在两方面：一是在主题树检索网络的系统中为用户组织、利用、评价信息提供了便利；二是由于与内容分类法相辅相成，互相补充，使得网络资源的组织与评价更为