

网上教学资源挖掘与文本自动分类系统

杨 晔

(广东商学院 教育技术中心, 广东 广州 510320)

摘要: 对数据挖掘技术在網上教学资源与知识分类系统中的应用进行了尝试性的研究, 并在广东商学院教学资源库建设中得到了成功应用.

关键词: 资源库; 数据挖掘; 知识分类; 贝叶斯概率

中图分类号: TP311. 13

文献标识码: A

文章编号: 1007-7162(2005)02-0079-04

网上教学资源库是一种基于 Internet 网络的自主学习和协作式教学系统, 它具有非常广泛的应用前景, 将人工智能与这种协作系统相结合, 将是它的未来发展方向^[1]. 教育技术专家桑新民教授在《现代教育技术学基础理论创新研究》中提出当代信息技术前沿跟踪与教育应用研究(尤其是教育人工智能研究)^[2]是我国教育技术领域的三大战略重点之一.

教育技术学的主要技术基础包括: 多媒体技术、数据库技术、通讯技术. 将这三大技术基础整合起来的是教育人工智能技术. 各种复杂技术要想在教育中广泛采用, 必须人性化、“傻瓜化”, 也就是智能化. 因此, 教育人工智能技术必将成为由教育技术学专业领衔, 广泛吸收多学科专家联合攻关的重大前沿课题^[3].

Internet 的迅猛发展使 Web 上的各种资源在几年间呈现爆炸式的增长. 互联网上容纳了海量的各种类型的原始信息资源, 包括文本信息、声音信息、图像信息、视频信息等等. 数以亿计 Web 网页中蕴涵了丰富的知识和教学资源. 如何在浩若烟海而又纷繁芜杂的 Web 网页中发现新资源并及时补充到资源库中, 已成为教育技术领域的一项极具挑战性的研究工作.

1 问题的提出

教学资源库的建设是高等教育信息化的基础. 它包括媒体素材库、试题库、案例库、课件库、资源索引库、网络课程库等. 作为数字化教学资源中心的网上教学资源库是知识经济时代开展创新教育、培养具备专业知识和信息素养的大学生的核心组成部分. 紧跟时代的发展和知识的不断更新是其重要的特点之一. 因此需要不断挖掘网上和网下的新知识、新资源, 并对挖掘的资源进行分类和整理, 及时向资源库中增加新的知识和内容.

面对网上的海量信息, 人们一般通过搜索引擎进行信息查询, 对网上信息进行人工选择和下载并加以分类和整理. 这种做法存在着许多弊端: 一是耗费大量的人力、物力和精力. 二是存在查询和分类结果一致性不高的问题. 即使分类人员的专业素质较高, 但不同的人来查询和分类, 其结果仍然不尽相同. 甚至同一个人在不同的时间做查询和分类, 也可能会有不同的结果.

收稿日期: 2003-12-10

作者简介: 杨晔(1964), 男, 高工, 主要研究方向为计算机网络、数据挖掘、办公自动化.

这种方法不利于资源库中知识的及时更新和资源库的长期发展.

为了解决上述问题, 我们有必要研究网上资源挖掘与文本自动分类系统, 提高资源库知识扩充和知识更新的速度和效率.

2 网上教学资源挖掘与文本分类系统

人工智能的一个重要支柱是数据挖掘技术(Data Mining). 数据挖掘一开始只是一种从大规模数据库或数据仓库中提取隐藏的预测性信息的科学方法. 它让人们有能力最终认识数据的真正价值, 即数据中潜在的可用信息和知识. 数据挖掘是数据库知识发现的核心步骤, 它研究的主要目标是发展有关的方法论、理论工具, 以支持从大量数据中提取有用的和让人们感兴趣的知识、模式和规则. 其主要研究内容包括基础理论、发现算法、数据仓库、可视化技术、定性定量互换模型、知识表示方法、知识分类算法、半结构化和非结构化数据中的知识发现以及网上数据挖掘等.

我们采用数据挖掘技术构建了一个具有网上资源挖掘和知识自动分类的教学资源库, 系统的框架如图 1 所示.

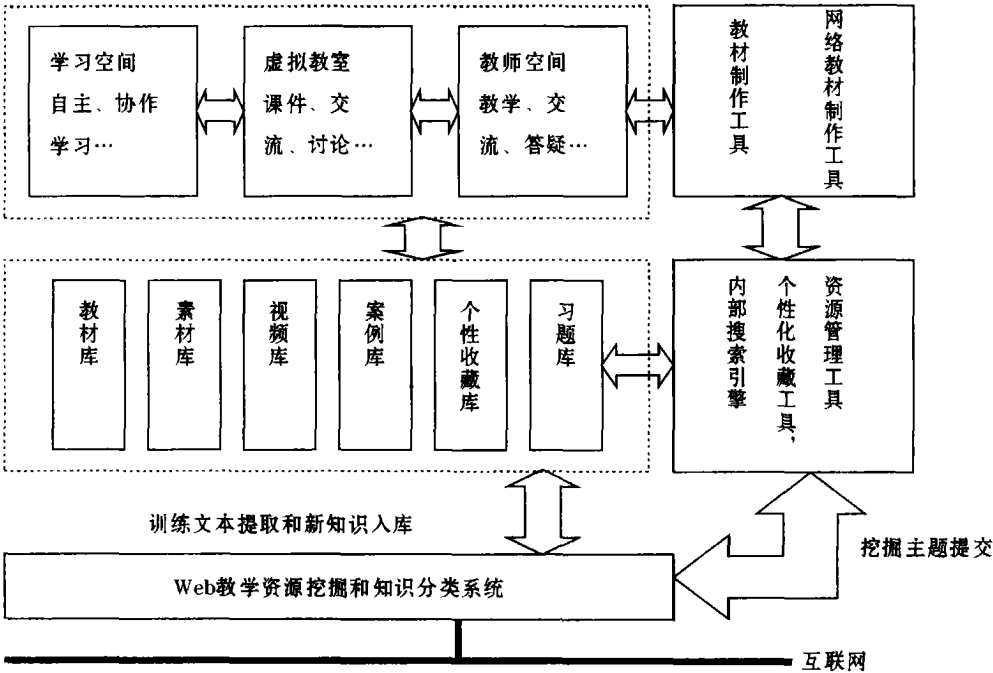


图 1 具有网上教学资源挖掘与文本自动分类的教学资源库

资源挖掘和知识自动分类系统由以下几个部分构成.

2.1 搜索引擎

从教师和系统管理员提交的资源挖掘主题中提取挖掘任务, 提交给搜索引擎通. 搜索引擎通过 Internet 发送蜘蛛(Spider) 机器人程序, 自动地从所爬行过的网页上抽取信息, 然后将网页连同其 URL 地址一起存入搜索引擎的索引数据库中.

2.2 网页信息分析

网上信息的数据结构跟数据库中存储的信息的数据结构有很大的不同, 并没有规定传递信息的具体格式, 这些信息的含义是由信息的发送者和接收者共同规定的. 因此这些 Web 网页的结构性一般都很差, 不能直接用于分类. 要解决这个问题, 必须对 Web 网页进行分析和信息提取.

2.3 中文网页的特征提取

1) 中文文本的自动分词

中文与英文的不同之处很多, 其中之一是中文句子中各个词条间没有固有的分隔. 首先需要对中文文本进行词条切分(简称分词), 采用基于词典的分词技术将常用词切分出来, 对于没有出现在词典中的未登陆词采用基于 n -gram 技术^[4,5]进行提取.

2) 文本特征选择

采用某种评估函数对每个文本特征独立进行打分, 把特征按分值高低排列, 通过设置阈值的办法选择预定数目的最佳特征作为文本特征子集^[6].

(1) 滤掉禁用词、停用词(例如: 的、地、得、如果、虽然等高频词和连词、虚词).

(2) 处理一词多义问题(同义关系、近义关系、从属关系、关联关系等).

(3) 给属性加权, 常用权重评估方法有:

a. 文本频数 DF;

b. 信息增益 Information Gain;

c. 互信息 MI;

d. 相对熵(Cross-Entropy);

e. χ^2 估计(CHI).

我们采用了互信息选择权重评估和特征词分布聚类方法^[7]抽取文本的特征词.

2.4 文本表示模型和新文本自动分类

常用的文本表示模型有布尔模型、向量空间模型 VSM、概率模型等. 概率模型的基本思想是利用词和文本的联合概率估计文本所属类别的概率, 概率模型的文本分类方法采用朴素贝叶斯分类算法:

假设文本是由各种词根据某种概率连续出现而形成的序列构成, 即文本中词的出现依赖于文本类别, 但不依赖于其它词及文本的长度, 也就是说词与词之间是独立的. 已知文本集包含 m 个文本类别变量 $C = \{C_1, \dots, C_j, \dots, C_m\}$, n 个文本变量 $D = \{d_1, \dots, d_i, \dots, d_n\}$, V 个特征词构成的特征词汇集 $W = \{w_1, \dots, w_k, \dots, w_v\}$, 一组带有先验的信息 $\theta = \{\theta_1, \dots, \theta_j, \dots, \theta_m\}$, 则文本中每个词 w_k 在文本 d_i 中出现的概率 $P(w_k | d_i)$ 用测试文本 d_i 中 w_k 出现的次数来表示:

$$P(w_k | d_i) = \frac{TF(w_k, d_i)}{\sum_{k=1}^V TF(w_k, d_i)} = \frac{TF(w_k, d_i)}{|d_i|}.$$

由资源库中提取的足够的训练文本, 从训练文本中可以得到 C_j 先验概率的值即: $P(C_j) = \frac{|C_j|}{|D|}$. 其中 $|C_j|$ 为训练集中某一类的文档数, $|D|$ 为训练集的总文本数. 还可以得到后验概率

$P(w_k | C_j)$ 的值: $P(w_k | C_j) = \frac{DF}{C_j}$, DF 为 C_j 类中出现 w_k 的次数, $|C_j|$ 为训练集中该类文本的总

数. 文本 d_i 属于类别 C_j 的后验概率为

$$P(C_j | d_i) = \frac{P(C_j) P(d_i | C_j)}{P(d_i)} = \frac{P(C_j) \prod_{k=1}^{|d_i|} P(w_{ik} | C_j)}{\sum_{r=1}^{|m|} P(C_r) \prod_{k=1}^{|d_i|} P(w_{ik} | C_r)},$$

对所有的 $i \neq j$, 求使 $P(C_i | d_i) > P(C_j | d_i)$ 的 C_i , 则类 C_i 为文本 d_i 的最终分类.

2.5 新资源的入库

将挖掘的网页(或纯文本)和它们的 URL 按照资源库的知识分类体系存入资源库中.

3 总结与展望

通过 Web 挖掘和机器学习技术挖掘 Web 上的教学资源, 将 Web 网页自动转换成纯文本格式, 对文本进行自动分类, 从而提高了系统的检索速度和检索精确度, 克服了人工查询和分类中信息检索不全面、更新速度慢的缺点. 我们将数据挖掘技术应用于网上教学资源库建设中, 这仅仅是一次成功的尝试, 今后还准备继续开展基于资源库日志挖掘的学生个性化学习系统的研究. 随着技术的不断成熟, 各种人工智能技术必将在教育资源库的建设中发挥重要作用.

参考文献:

- [1] 何克抗. 基于 Internet 的教育网络与 21 世纪的教育革新[EB/OL]. <http://202.112.88.32/论著选摘/何克抗/JYU-Internet.htm>, 2003-04-03.
- [2] 桑新民. 现代教育技术学基础理论创新研究[J]. 中国电化教育, 2003, (9): 12.
- [3] 谢百治. 美国教育技术简介[J]. 电化教育研究, 2001, (9): 55.
- [4] 何浩, 杨海棠. 一种基于 N-Gram 技术的中文文献自动分类方法[J]. 情报学报, 2002, 21(4): 421.
- [5] 中国科学院. 计算机汉语词法分析系统 ICTCLAS[EB/OL]. http://www.nlp.org.cn/project/project.php?proj_id=6, 2002-08-16.
- [6] 朱明, 王军, 王俊普. Web 网页识别中的特征选择问题研究[J]. 计算机工程, 2000, (8): 35.
- [7] L Douglas Baker. Distributional Clustering of Words for Text Classification [EB/OL]. <http://citeseer.ist.psu.edu/65816.html>, 2003-04-03.

Web Instructional Resources Mining and Text Categorization System

YANG Ye

(Educational Technology Center, Guangdong Commercial College, Guangzhou 510320, China)

Abstract: This paper studies the application of data mining in a web instructional resource base. A system of web instructional resources mining and text categorization has been developed and applied successfully in the instructional resource base of Guangdong Commercial College.

Key words: instructional resource base; data mining; text categorization; NaiveBayesian probability