●罗 琳 陈 远

知识挖掘与数字图书馆个性化服务*

摘 要 数字图书馆要实现个性化服务,应当:(1)收集与用户有关的信息,建立用户信息库。(2)利用知识挖掘技术,分析用户信息库。(3)跟踪本地信息库和网络信息,收集用户所需信息。(4)分析用户反馈信息。(5)针对不同类别的用户,提供交流平台。参考文献4。

关键词 数字图书馆 知识挖掘 个性化服务 用户服务
分类号 G252

ABSTRACT To realize personalized services of digital libraries, we should do the following things: 1) collecting user-related information and creating user information database, 2) analyzing user information database by knowledge mining techniques, 3) tracking local information database and network information to collect information required by users, 4) analyzing user's feedback information, 5) providing exchange platform different users. 4 refs.

KEY WORDS Digital library. Knowledge mining. Personalized services. User services.
CLASS NUMBER G252

1 数字图书馆个性化服务

数字图书馆用户的需求不仅是简单查询。他们希望通过对数据的较高层次的处理与分析,得到关于数据总体特征和对发展趋势的预测。数字图书馆应当从多种媒体表示的信息中,以面向"个人"的方式来挖掘知识。

数字图书馆的信息容量十分巨大,覆盖从图像、图形、文字、语音到电子文档等各种媒体类型,以格式化的关系数据库和非格式化的电子文档、HTML网页等为其存储方式。只有突破传统检索服务的樊篱,开创知识挖掘服务的新境,才能适应用户需求。但信息量的无限性,加大了处理难度。

电子文档、网上超文本信息和数据库里的大字段信息中包含大量的冗余信息和信息垃圾。提取真正有用的信息,并以可读、精炼、概括的形式,将处理结果提交给用户,提高工作效率和资源利用率,是数字图书馆建设中要考虑的问题之一。

对同样的数据,不同领域、不同工作岗位上的用户其期望值不同。数字图书馆要提供一种因人而异的、有针对性的主动服务。个性化服务目前主要表现在:(1)个性化链接。个性化链接用于收集和组织由用户定义的资源。其基本设计思想是利用文件夹包含个人选择的链接地址,链接用户选择的资源。(2)个性化更新。确定用户感兴趣的领域,在与这些领域相关的资源变化时,系统及时通知用户。(3)虚拟社团。这种模式下,由数字图书馆提供一个统一平台,实现不同类型用户的信息共享和互动。知识挖掘技术有助于数字图书馆开展个性化服务。

2 知识挖掘的概念

知识挖掘是所谓"数据挖掘"的一种更广义的说法,其定义几经变动。最新的描述性定义是:按照某种既定目标,对大量数据进行分析和探索,从中识别出有效的、新颖的、潜在有用的知识,以最终可理解的模式显示的一系列处理过程。它涉及到机器学习、模式识别、统计学、数据库、联机分析、模糊逻辑、人工神经网络、不确定推理等多种学科知识。从数据库中发现出来的知识可以用在信息管理、过程控制、科学研究、决策支持等许多方面。

2.1 知识挖掘步骤

(1)确定应用领域:包括此领域的基本知识和目标。

(2)建立目标数据集:选择一个数据集或在多数据集的子集上聚焦。

(3)数据预处理:在大数据集中,根据需求,利用数据净化和整合技术,选择与任务相关数据,在不降低其准确度的状况下减少处理数据量。

(4)数据转换:找到数据的特征进行编码,减少有效变量的数目。

(5)数据挖掘:根据数据和所要发现知识的种类来确定相应的挖掘算法。

* 本文系国家社科基金项目(项目编号00BTQ004)研究成果。

(6)数据评价:将挖掘出的知识和数据以各种可视化方式显示,并将其以图形、文本等方式存储在库中,以便对它们进一步挖掘,直至满意为止。

2.2 知识挖掘关键技术和方法

2.2.1 仿生技术

(1)人工神经网络技术。人工神经网络建立在可以自学习的数学模型的基础上,对大量复杂的数据进行模式抽取及趋势分析。其比较典型的学习方法是回溯法。它通过将输出结果的一些已知值进行一系列比较,不断调整加权值,得到一个新的输出值,再经过不断的学习过程,最后,该神经网络得到一个稳定的结果。

(2)遗传学算法。这是模拟生物进化过程的算法。首先对求解的问题进行编码,再利用基因复制、突变等方法产生新的个体,直至求得最佳个体。

2.2.2 决策树技术

它根据信息论原理,寻找数据库中具有最大信息量的字段,建立决策树的一个结点,再根据字段的不同取值建立树的分枝;在每个分枝中集中重复建立树的下层结点和分枝的过程,即可建立决策树。该类方法的实用效果好,影响较大。

2.2.3 统计学方法

利用统计学原理对数据库中的信息进行分析,常用的有因子分析、相关分析、多元回归分析、最小二乘法分析等。

2.2.4 集合论方法

(1)覆盖正例排斥反例方法:它是利用覆盖所有正例,排斥所有反例的思想来寻找规则。

(2)概念树方法。数据库中记录的属性字段按归类方式进行合并,建立起来的层次结构称为概念树。利用概念树提升的方法,可以大大浓缩数据库中的记录。对多个属性字段的概念树提升,将得到高度概括的知识基表,再将它转换成规则。

(3)粗集(Rough Set)方法。近年来,波兰华沙理工大学 Z. Pawlak 教授等一批科学家提出了用粗集理论,研究不完整数据、不精确知识的表达、学习、归纳等方法。粗集理论出发点在于认为知识是基于人们对研究对象分类的能力。根据当前已有的关于给定问题的知识,将问题论域进行划分,然后对划分后的每一部分确定其对某一概念的支持程度,分为:肯定支持此概念、肯定不支持此概念和可能支持此概念。粗集理论应用的主要思想,是在保持分类能力不变的前提下,通过知识的简约,导出概念的简练分类知识。

2.2.5 其他技术

(1)规则推理方法。该方法在数据集中对目标数据进行规则的搜集,建立起数据库的规则分析网,主要包括:关联规则分析、多层次规则分析以及差别规则分析等。

(2)聚类技术。在分类的基础上,根据对象的特征,从宏观上控制同一类别的数据。

(3)可视化技术。可视化数据分析技术拓宽了传统的图表功能,从而可以更清楚地分析数据。

3 知识挖掘在数字图书馆个性化服务中的应用

(1)收集与用户有关的信息,建立用户信息库。用户是数字图书馆的重要资源。一个信息完整的用户信息库,能保证在充分挖掘的基础上,了解用户的普遍性需求与特殊性需求,从而开展针对性服务。用户信息库应尽可能包括用户的基本情况,如用户姓名、身份、单位等,还应该包括用户的兴趣、爱好、研究领域、知识结构、习惯行为方式等。得到用户信息主要有以下方式。

第一,用户自行登记:数字图书馆可以在主页上提供个性化服务的注册入口。在进入页面,设计一个表单,内容包括用户的基本信息和特殊信息,如感兴趣的专业、方向,想获得的信息以及获取方式和时间间隔、联络方式等。比如武汉大学图书馆提供的 Email-Alert 服务,就可以根据用户的兴趣将最新信息发送到用户的邮件信箱。

第二,跟踪日志:数字图书馆可以在用户从进入该图书馆网站开始,对用户的行为进行跟踪,产生日志文件,通过对日志文件的分析,了解用户使用该图书馆的主要目的和其需求。并将该日志文件并入用户信息库。这里涉及到用户识别和用户浏览页面的识别。目前一般通过3种方式来识别用户。第一种是使用 cookies 的方式。它是浏览器提供的信息记录区,任何 Web 服务器都可以利用这个记录区来记录信息,并可在其中放入 ID 号来识别用户。但并不是每一种浏览器都具有该功能,因而应注意写入时的时间设定并注明域名,以保证只有下次进入该域名时才能读取这个 cookies。第二种方法是利用 URL。用户请求的 URL 被放上一个标识,只有当用户不断请求时才能识别用户。第三种方法是使用用户的 IP 地址。由于许多 IP 地址是动态分配的,而且可能许多用户共用一台计算机,这样一个 IP 地址可能代表许多毫无关系的用户。一般采用一些启发式的规则来帮助识别用户:如果 IP 地址相同,但代理日志中表明用户的浏览器或者操作系统变了,则认为不同的代理表示不同的用户;将访问日志、引用日志和站点拓扑结构结合,构造用户的浏览路径;如果当前请求的页面同用户已经浏览的页面间没有链接关系,则认为存在 IP 地址相同的多个用户。对于用户浏览页面的识别,主要是对用户行为的识别,了解他所浏览的页面时间、类型以及该页面的元数据等。

第三,系统导入:可以通过联合图书馆方式将其他数字图书馆的用户信息库导入,也可以与其他网站(如科研所)合作,导入其用户数据。

(2)利用知识挖掘技术,分析用户信息库。为了开展个性化服务,必须了解用户,进行以下分析。

第一,用户分类。可以通过用户所处地理位置、年龄、从事职业、所在行业、利用图书馆信息服务程度、用户对图书馆提供服务手段掌握的熟练程度等方面将用户进行分类。

第二,用户行为分析。主要可以通过对用户利用图书馆的情况和对数字图书馆网站访问内容分析。关于前者,可以通过用户业务处理中的数据得知该用户的兴趣和爱好。这些数据来源主要是:用户的借阅信息和咨询记录等。对于后者,则是通过用许多曲线图解法对所浏览的页面路径进行分析。最典型的就是用曲线来表示站点的物理布局,曲线中的点代表 Web 页面,连线代表页面之间的超链接。基于网页的类型可以形成其他的一些图形,利用图形的边代表页面之间的相似性,或者在边上给出使用该超链接的人数。可以从图形的物理布局中找到用户的浏览模式,确定网站中频繁访问的路径。

第三,通过对用户分类、聚类和时间序列模式分析,抽象出每类用户的普遍性需求和个性化需求,并建立一系列关联规则。如某图书馆新引进一批国外信息管理专业学术论文,则 80%以上信息管理专业学生需要,95%以上信息管理专业教师需要,但生产人员需要率可能不到 1%,这样就可以将论文信息传递给需要率较高的相关用户,提供他们使用,对于需要率不足 20%的根本不需要通知。还可以通过对网站中频繁访问的路径的确定和发现用户的引用页面和服务器上多种页面之间的联系,分析特定用户的访问模式和浏览行为,得出其兴趣爱好,然后动态调整、定制网站中页面链接的次序以及网站页面内容,向用户推荐他们可能感兴趣的内容,进而创建智能 Web 站点,实现个性化链接。或者辅助改进分布式网络系统的设计性能,如在有高度相关的站点间提供快速有效的访问通道,帮助更好地组织设计 Web 主页。

(3)根据用户需求,主动跟踪本地信息库和网络相关信息,收集用户所需信息。为了提高信息的准确度,还应对所收集到的信息进行相关性分析。此相关性分析,可以根据用户提供的检索词,确定所检索到的信息与该检索词的相关度。目前最常用的相关度分析,是通过限制域(如标题、正文、文摘甚至是文献类型)或词频控制。

(4)信息推送与反馈。数字图书馆可以利用智能推送技术将用户所需信息推送到用户的计算机、电子信箱,甚至手机、PDA 上。还可以提供标题、简介或 URL 等方式供用户选择信息内容。为了提高服务质量,可以要求用户对所提供的信息作出反馈(如该类信息是否继续提供、相关信息是否需要、取消定制等)。并将用户的反馈信息再次分析,调整和完善用户信息库。

(5)数字图书馆除主动信息服务外,还应该根据用户分析结果,针对不同类别的用户,提供交流平台,建立虚拟社区,实现信息共享。

4 应注意问题

(1)挖掘算法的效率和可扩放性。目前数据库数量大,维数高,使得挖掘的搜索空间增大,发现知识的盲目性提高。如何充分利用领域的知识,剔除与发现任务无关的数据,有效地降低问题的维数,设计出高效率的知识发现算法是下一步发展的重点。需要解决的问题有:开发适宜于大数据集的分类、聚类、相关性分析、变异性探测的有效算法;开发在混合数据集(类型变量与数值变量混合)上有效的数据抽样、数据缩减及维数缩减方法;开发既考虑数据先验知识使搜索简化,又考虑成本效益且对不确定性和丢失数据具有稳健性的数据挖掘算法;开发能抽取样品间复杂关系且考虑其结构(如稀疏关系)的数据挖掘搜索新算法。

(2)用户数据的质量问题。用户数据残缺不利于全面了解用户信息,会对用户的个性分析产生误差,影响服务质量。这主要包括两方面残缺:用户数据的残缺和用户信息库设计上的残缺。对于前者,我们应该通过对用户行为的分析,补全相关信息;或通过信息反馈方式,不断了解用户的真正需求,将其需求数据化或潜在需求显化,从而完成用户信息的获取。对于后者,可能由于初始设计过程中将应该提取的用户特征忽略,必须通过不断实践,将数据库设计完善,将所需要的用户特征提取。

(3)服务质量控制。应该从方便性、及时性、交互性和指导性等方面,提高服务质量。所谓方便性,指开发更多个性化服务手段,为用户提供更多的选择。所谓及时性,指要根据用户行为方式的变化,随时调整用户信息库,提供及时服务。所谓交互性,指要不断地与用户交流,了解用户的需求以及对数字图书馆的信息提供的反馈意见,及时调整策略。所谓指导性,指在服务过程中应该不断指导用户,提高使用技能,提高其信息素质。

参考文献

1 吕安民等.数据挖掘和知识发现的技术方法.测绘科学,2000(4)
2 韩惠琴,刘柏嵩,董其军.知识发现在数字图书馆中的应用.大学图书馆学报,2001(1)
3 史田华.论 Internet 知识挖掘.图书情报知识,2002(3)
4 Jiawei Han 等.数据挖掘—概念与技术.北京:高等教育出版社,2001

罗 琳 武汉大学信息管理学院教师、博士。通讯地址:武汉大学信息管理学院。邮编 430072。

陈 远 武汉大学信息管理学院副教授。通讯地址同上。

(来稿时间:2003-09-25)