Web数据挖掘综述

麦晓冬 余海冰
(广东轻工职业技术学院 510300)

摘 要:如何在Web这个全球最大的数据集合中发现有用的信息成为了数据挖掘研究的热点。文章首先分析了Web数据挖掘所面临的问题,然后概述了Web数据挖掘以及它的几个分类,最后简单阐述了Web挖掘的应用前景以及在Web2.0到来之时,Web数据挖掘所面临的机遇与挑战。

关键词:Web数据挖掘 内容挖掘 结构挖掘 使用挖掘 用户性质挖掘 Web2.0
中图分类号:TP311 文献标识码:A 文章编号:1672-3791(2007)02(a)-0012-02

Web目前已成为一个巨大的、分布广泛的和全球性的信息服务中心,它涉及新闻、金融管理、教育、广告、娱乐、电子商务和许多其它信息服务。Google、Baidu等的搜索引擎的出现,正是满足了人们从这样一个信息服务中心中寻找有用的知识的需要。虽然当前的搜索技术能在一定程度上帮助我们从Web中获取有用的知识,但是精度不够,不能满足实际需要,更谈不上挖掘出蕴含在Web数据背后的知识。如何从Web上海量的数据中找到真正有用的信息成为人们关注的焦点,Web数据挖掘技术也正是伴随着这种需求从研究走向应用。

1 Web数据挖掘概述

Web数据挖掘是从数据挖掘发展而来,是数据挖掘技术在Web技术中的应用。Web数据挖掘综合运用了统计学、计算机网络、数据库与数据仓库、可视化等众多领域的技术。通过从Web上的资源中抽取信息来提高Web技术的利用效率,也就是从Web文档结构和使用的集合中发现和分析潜在的、有用的模式或信息。

一般的,Web数据挖掘的原理可用图1所示的过程表示。目标数据集就是根据实际需要所提取的Web数据;预处理就是从目标数据集中清除掉明显错误的数据和冗余的数据,并进一步将数据转换为适用于数据挖掘的形式;模式发现是选择合适的算法来处理经预处理的数据,并最终发现用户的访问数据;模式分析是对发现的模式进行分析评估,必要时需要返回前面处理中的某些步骤反复提取;最后的信息将通过可视化、联机分析等技术处理,使之能以易于理解和接受的方式显现出来。[1]

本文主要论述Web数据挖掘以及Web2.0的出现给Web数据挖掘带来的影响。

2 Web数据挖掘面临的问题

由于Web上信息的特点,对Web进行有效的信息挖掘,发现有用的知识信息具有很大的挑战,同时也面临很多的问题。

2.1 异构数据库环境 从数据库的研究角度出发,Web中的信息可以看作巨大的数据库:每一个站点就是一个数据源。而且站点之间的信息和组织不同,因而构成的是巨大的异构数据库环境。如果要利用这些数据进行挖掘,必须研究站点之间异构数据的集成问题。只有集成这些站点的数据,为用户提供统一的视图,才有可能从巨大的数据资源中获取所需的内容。还要解决Web上的数据查询问题,因为如果所需的数据不能很有效得到,则分析、集成并处理这些数据就无从谈起。

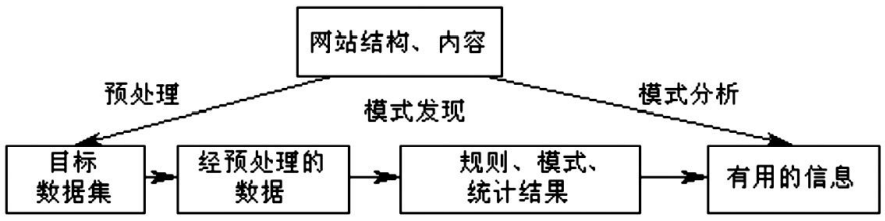

图1 Web数据挖掘原理

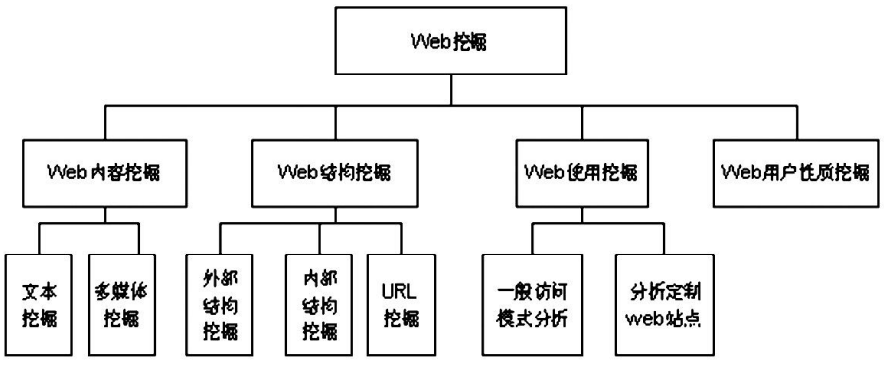

图2 Web数据挖掘的分类

2.2 半结构化的数据结构

半结构化是Web上数据的最大特点。传统数据库都有一定的数据模型,可以根据该模型具体描述特定的数据;而Web上的数据非常复杂,没有特定的模型描述。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,即半结构化数据。

2.3 解决半结构化的数据源问题

Web数据挖掘技术首先要解决半结构化数据源模型及其查询和集成问题,解决这个问题必须要有模型清晰地描述Web上的数据,查询一个半结构化的数据模型是关键所在。除定义这个模型外,还需要一种自动地从现在数据中抽取半结构化模型的技术。面向Web的数据挖掘必须以半结构化模型和半结构化数据模型抽取技术为前提。[2]

3 Web数据挖掘的分类

一般来说,众多Web数据挖掘相关文献上将Web数据挖掘分为三类:Web内容挖掘、Web结构挖掘、Web使用挖掘。但是随着Web2.0的出现,Web数据挖掘多出了一个分类——Web用户性质挖掘。Web数据挖掘的分类如图2所示。

3.1 Web内容挖掘

Web内容挖掘主要包括文本挖掘和多媒体挖掘两类,其对象包括文本、图像、音频、视频、多媒体和其他各种类型的数据。这些数据一般由非结构化的数据(如文本)、半结构化的数据(如HTML文档)和结构化的数据(如表格)构成。对非结构化文本进行的Web挖掘,称为文本数据挖掘或文本挖掘,是Web挖掘中比较重要的技术领域。Web文本挖掘的一般处理过程如图3所示。

目前,关于Web内容挖掘的研究大体以Web文本内容挖掘为主。Web内容挖掘一般从资源查找和数据库两个不同的方面进行研究。

(1)从资源查找的方面来看,Web内容挖掘的任务是从用户的角度出发,着重提高信息质量和帮助用户过滤信息。主要是对非结构化文档和半结构化文档的挖掘;

(2)从数据库的观点进行Web内容挖掘主要是试图建立Web站点的数据模型并加以集成,以支持复杂查询,而不只是简单的基于关键词的搜索。这要通过找到Web文档的模式、建立Web知识库来实现。

对文本数据进行挖掘的文档分类和模型质量评价方法与传统的数据挖掘方法相类似，分类算法主要应用朴素贝叶斯(Naive Bayes Classifier)。对模型的质量评价主要有分类的正确率(Classification Accuracy)、准确率(Precision)和信息估值(Information Score)。

Web 多媒体数据挖掘从多媒体数据库中提取隐藏的知识、多媒体数据关联、或者是其他没有直接储存在多媒体数据库中的模式。多媒体挖掘首先进行特征提取，然后再应用传统的数据挖掘方法进行进一步的信息挖掘。对网页中的多媒体数据进行特征的提取，应充分利用 HTML 的标签信息。[3]

3.2 Web 结构挖掘

Web 结构挖掘是指挖掘 Web 潜在链接结构模式，即通过分析页面链接和被链接数量以及对象来建立 Web 自身的链接结构模式。Web 数据不同于文本或者数据库，有用的知识不仅在Web 页面的内容中存在，而且也在 Web 页面间的链接结构和 Web 页面内部结构中包含。所以，Web 结构挖掘可以分为外部结构挖掘、内部结构挖掘以及 URL 挖掘。

Web 结构挖掘的基本思想是将 Web 看作一个有向图或者无向图的形式，把 Web 页面抽象作为图的顶点，而页面间的超链接就是图的边。然后利用图论对 Web 的拓扑结构进行分析研究。常见的算法有 PageRank、HITS (Hypertext Induced Topic Search)、二次方程推断法 (Quadratic Extrapolation)、分块矩阵排序算法(Block Rank Algorithm)、发现虚拟社区 (Cyber-community) 的算法、发现相似页面的算法等。[4]

Web 结构挖掘的算法一般可分为查询相关算法和查询无关算法两类：

(1)查询相关算法需要为每一个查询进行一次超链分析从而进行一次值的指派；

(2)查询独立算法则为每个文档仅进行一次值的指派，对所有的查询都使用此值。

3.3 Web 使用挖掘

Web 使用挖掘是从用户存取模式中获取有价值的信息，即通过分析 Web 日志数据及相关数据，来发现访问者访问 Web 页面的模式，分析日志记录中的规律，从而识别访问者的兴趣、频率、满意度，可以发现潜在用户，增强站点的服务竞争力。

Web 内容挖掘、Web 结构挖掘的对象是 Web 上的原始数据，而 Web 使用记录挖掘则面对的是在用户和网络交互的过程中抽取出来的第二手数据。这些数据包括：网络服务器访问记录、代理服务器日志记录、浏览器端日志记录、用户简介、注册信息，用户对话或交易信息，cookie 中的信息、用户查询、等一切用户与站点之间可能的交互记录。一般可分为一般访问模式分析以及分析特制 Web 站点。[5]

Web 使用挖掘的基本流程包括数据预处理、模式识别和模式分析。其基本流程如图 4 所示。

数据预处理：将各种收集到的“第二手数据”进行数据清洗、用户识别、会话识别、路径补充、事务识别，进行预处理主要是为了后续的模式识别提供结构化的、可靠的、整合的数据。

模式识别：主要采用数据挖掘领域的一些技术和算法，对 Web 使用模式进行挖掘，如路径分析、关联规则、聚类和分类等。

模式分析：通过模式识别算法找到的模式集合中筛选出有意义的模式、规则。需要一些分析工具的辅助，如果没有合适的技术和工具来帮助分析人员理解，挖掘出来的模式将得不到很好的利用。常用的模式分析技术有:知识查询、可视化技术 (Visualization)、联机分析处理(OLAP)等。

3.4 Web 用户性质挖掘

Web2.0 是从 2005 年直到现在一直都很流行的名词。Web2.0是以Flickr、Craigslist、Linkedin、Tribes、Ryze、Friendster、Del.icio.us、43Things.com 等网站为代表，以 Blog (博客 / 网志)、TAG (网页书签)、SNS (社会网络)、RSS (站点摘要)、wiki (百科全书) 等应用为核心，依据六度分隔、xml、Ajax 等新理论和技术实现的互联网新一代模式。[6]Web2.0 时代的显著特征是个性化、互动性、大众化和去中心，旨在给用户提供更人性化的服务，同时不再像 Web1.0 时代用户只能被动的接受各网站“填鸭”式的信息轰炸。在 Web2.0 时代，每个普通用户既是信息的获取者，也是信息的提供者。面对 Web2.0 的诞生，Web 数据挖掘技术又面临着新的挑战。[7]

如果说 Web 使用挖掘是通过挖掘网站访问者在网站上留下的痕迹来获取有用的信息，那么 Web 用户性质挖掘则是要去 Web 用户的老巢去探寻究竟。在 Web2.0 时代，网络彻底个人化了，Web 用户可以用自己的方式、喜好来个性化定制自己的互联网。Web2.0 赋予 Web 用户最大的自由度，同时给予有心商家有待发掘的高含金量信息数据。通过对 Web 用户自建的 Blog、RSS 等 Web2.0 功能模块下客户信息的统计分析，能够帮助运营商以较低成本获得准确度较高的客户兴趣倾向、个性化需求以及新业务发展趋势等信息。有关 Web2.0 下的数据挖掘正在进一步研究中。

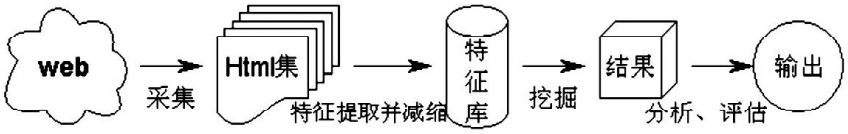

图 3 Web 文本挖掘的一般处理过程

图 4 Web 使用挖掘基本流程

4 结语

近年来，随着 Web 技术的发展，Web 挖掘的形式和研究方向层出不穷。Web 使用挖掘在电子商务相关方向有着较好的发展；在搜索引擎的研究方面，Web 结构挖掘的研究也日趋成熟；基于文本的内容挖掘已经有较深的研究，多媒体挖掘方面也逐渐成为新的研究热点。而 Web2.0 的出现给 Web 数据挖掘提出了新的要求，基于 Web2.0 的数据挖掘目前还处于起步阶段，它必将成为 Web 数据挖掘中很重要的一个研究领域。

参考文献

[1] 马保国,侯存军,王文丰等.Web 数据挖掘技术与应用 [J]. 计算机与数字工程,2006,6:20-22.

[2] 何鲲,朱方洲.基于web的数据挖掘方法的研究及实现[J].合肥学院学报,2005.6(15):24-27.

[3] 陈二忠,姜丽华.基于 Web 的数据挖掘技术[J].微机发展,2003,12:61-64

[4] 张小松,窦炳琳.Web 挖掘研究[J]. 唐山学院学报,2003,12:80-84.

[5] 曼丽春,朱宏,杨全胜.Web 数据挖掘研究与探讨[J].现代电子技术,2005.8:3-6

[6] http://homepage.yesky.com/300/2295800.shtml.全面解读 web2.0.截至到2006,12,10.

[7] 高祥华.Web2.0 中的技术及应用[J] .中国科技信息.2006,(13):127-128.