Web 挖掘技术探讨

文⊙ 张春明（辽宁对外经贸学院）

摘要:随着Internet的迅猛发展，web挖掘逐渐成为数据挖掘的热点，但是因为Web自身的特点：多数据源，数据结构的半结构化，及动态性等种种，Web挖掘又是一个难点。本文从web挖掘的内涵入手简要介绍了web挖掘的目标、

关键词:Web挖掘；数据挖掘；聚类

因特网使人们获得信息的流行和重要手段，其发展带动了Web的发展，同时在电子商务的发展历程中，不同模式的电子商务网站应运而生，发展到现阶段，传统企业的加入使得电子商务发展到一个新的阶段。随着企业网站的规模加大，和复杂度的增强，人们对站点的设计和功能提出了更高的要求。要求Web具有智能性，能快速、准确地找到用户所需信息，能为不同用户提供不同的服务；能为用户提供产品营销策略信息等。在当前的信息分析技术中，Web挖掘是最具有应用前景的一种技术、分类。做到对web挖掘有个系统的介绍。

一、web 挖掘内涵

Web是一个巨大的、开放性、动态性和广泛分布、高度异构、半结构化、超文本、相互联系并且不断进化的信息仓库。它也是一个巨大的文档累积的集合，包括超链接信息、访问及使用信息，资源分布分散，这就导致信息获取的困难。Web数据挖掘技术的诞生为这一领域的知识发现带来了生机，以便人们可以从 Web海量的数据中自动地、智能地抽取隐藏在这些数据中的知识。

Web挖掘就是利用数据挖掘技术的思想和方法从Web访问日志中提取有用的模式，这些模式能够揭示站点访问者的有趣特性。

二、Web数据挖掘的目标

（一）提高服务器性能

对一个网站来说，需要运用最少的带宽和服务器资源，为更多的客户提供更快捷的服务。而用户对Web站点的满意度，主要以访问速度来衡量。从用户角度来说，Web站点只有“快”和“慢”之分；用户往往并不要求实现大容量数据传输，而是希望网站在保证性能的同时，能够容纳更多的访问者。Web用户所关心的问题的实质是访问时间。而对于网站运营方来希望通过web挖掘来提高服务器的性能，达到用户希望的访问速度。

（二）改善网站导航

用户访问网站的另一个评价标准是网站是否易用，方便易懂的网站导航使得网站在用户心目中的信用和地位都会有所上升。

（三）改善网站应用的系统设计

网站的设计归根结底是为了在提供给用户需求的信息和服务的同时，占有客户的眼球，好的网站应用坚实的先入为主的网络经济体现得更加明显。

（四）为电子商务寻找目标用户

web挖掘可以将用户的访问习惯挖掘出来，方便对用户的分类管理，和有目标的进行信息推送。

（五）发现潜藏的客户群

对于网站长期目标是维持现有客户挖掘潜在客户，达到客户群体的价值最大，将客户信心转化为客户价值。

三、Web挖掘研究分类

Web页面是互联网上存储和发布信息最普遍的载体，是世界上最大的信息仓库之一。Web上存储的信息量巨大而且缺乏结构化组织的规整性，人们访问Web留下的日志也是海量数据。近几年数据挖掘技术不断的发展完善，为Web信息的处理和有效使用提供了有效的工具。Web挖掘已经成为数据挖掘技术一个重要的应用领域。现今最流行的对Web数据挖掘的分类是根据挖掘的对象将其分为三类：

(一)基于Web内容(Content)的挖掘

Web内容挖掘是指对Web页面内容进行挖掘，从文本，图像，音频，视频，动画等各种形式的网络资源中发现所需的特定化信息，以实现Web资源的自动检索，提高Web数据的利用效率。Web数据分布范围很广，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及无结构的自由文本数据信息，有在FTP上的，在Gopher中的，在数字图书馆中的，还有企业自己Web网站上的，也有隐形的私人数据和动态查询的结果等。

(二)Web结构(Structure)挖掘

Web结构挖掘是挖掘Web潜在的链接结构模式，找到隐藏在一个个页面之后的链接结构模型，该模型可用于网页重新分类，也可以用于寻找相似的网站，并由此获得有关不同网页间相似度及关联度的信息。这有助于用户找到指向相关主题的权威站点。

(三)基于Web使用记录(Usage)的挖掘

Web使用挖掘是挖掘Web服务器日志获取的信息来预测用户浏览行为的技术，指从用户的访问日志中挖掘用户的访问模式，为网站经营管理和结构调整提供决策支持，为企业发现新市场机会，进行市场决策，提高通过网站施行的营销效果，以及为企业进行战略决策提供有价值的潜在的信息。

四、web数据挖掘的特点

Web数据有其自身的特点：

（一）对有效的数据仓库和数据挖掘而言，Web似乎太庞大了

Web的数据量目前以兆兆字节(terabytes）计算，而且仍然在迅速地增长。许多机构和社团都在把各自大量的可访问信息置于网上。这使得几乎不可能去构造一个数据仓库来复制、存储或集成Web上的所有数据。

(二)Web页面的复杂性高于任何传统的文本文档

Web页面缺乏统一的结构，它包含了远比任何一组书籍或其它文本文档多得多的风格和内容。

(三)Web是一个动态性极强的信息源

Web不仅以极快的速度增长，而且其信息还在不断地发生着更新。新闻、股票市场、公司广告和Web服务中心都在不断地更新着各自的页面，Web日志更是每秒钟都会记录下大量的访问信息。

(四)Web面对的是一个广泛的用户群体

目前因特网上连接有约5千万台工作站，其用户群仍在不断地扩展当中。各个用户可以有不同的背景、兴趣和使用目的。Web上的大量信息相对于多数用户是无用的。用户只关心Web上的很小一部分信息，其余信息对用户来说是不感兴趣的，反而会淹没其所希望得到的搜索结果。

五、如何实现Web挖掘

Web挖掘发展自数据挖掘。数据挖掘方法通常可以分为两类：一类是建立在统计模型的基础上，采用的技术有决策树、分类、聚类、关联规则等；另一类是建立一种以机器学习为主的人工智能模型，采用的方法有神经网络、自然法则计算方法等。

（一）Web内容挖掘实现技术

Web上的内容挖掘多为基于文本信息的挖掘，它和通常的平面文本挖掘的功能和方法比较类似。利用Web文档中部分标记，如Title、Head等包含的额外信息，可以提高Web文本挖掘的性能。

文本总结。文本总结是指从文档中抽取关键信息，用简洁的形式对文档内容进行摘要或解释。其目的是对文本信息进行浓缩，给出它的紧凑描述。这样，用户不需要浏览全文就可以了解文档或文档集合的总体内容。

文本分类。分类是在已有数据的基础上学会一个分类函数或构造出一个分类模型，即通常所说的分类器。

文本聚类。文本聚类把一组文档按照相似性归成若干类别。方法大致可分为层次凝聚法和平面划分法两种类型。

关联规则。发现关联规则的算法通常要经过以下三个步骤：连接数据，作数据准备；给定最小支持度和最小可信度，利用数据挖掘工具提供的算法发现关联规则；可视化显示、理解、评估关联规则。

（二）Web使用记录挖掘实现技术

在挖掘Web用户使用记录时描述用户访问的数据包括：IP地址、参考页面、访问日期和时间、用户Web站点及配置信息。

发现用户使用记录信息的方法有两种。一种方法是通过对日志文件进行分析，包含两种方式：一是先进行预处理，即将日志数据映射为关系表并采用相应的数据挖掘技术来访问日志数据；二是直接访问日志数据以获取用户的导航信息。另一种方法是通过对用户点击事件的搜集和分析发现用户导航行为。

参考文献：

[1]毛国君,段立娟,王实,白云.数据挖掘原理与算法.[Z]北京:清华大学出版社,2005-7

[2]范明,范宏建.数据挖掘导论.北京:人民邮电出版社,2006-5

[3]顾晓燕,关于挖掘技术的研究,电脑知识与技术,2005

[4]薛鸿民,数据挖掘技术研究现,代电子技术,2006