Web 数据挖掘技术应用

郑泠
杭州师范学院钱江学院电气系 电子工程 031
310018

[摘　要] Web 数据挖掘是目前信息技术中的研究热点，它是现代科学技术相互渗透融合的必然结果。文章首先介绍了 web 数据挖掘的含义，重点讨论了 web 数据挖掘的类型以各种类型的 web 数据挖掘的基本过程以及它们所使用的一些相关技术及应用，并对数据挖掘的发展前景和方向进行了展望。

[关键词] 数据挖掘；Web 数据挖掘；相关技术；应用

引言

Internet 给人类带来了巨大的变革，随着 Internet 的进一步发展和完善，各种基于 Internet 的应用业务也如雨后春笋般的发展起来，例如网上商店、网上银行、远程教育、远程医疗等。毫无疑问未来的商战战场将是 Internet 。同时，我们也应该看到 Internet 在给我们带来机遇的同时也带来了挑战，它使得 WWW 上的一些主要工作，例如 Web 站点设计、Web 服务、Web 服务设计、Web 站点的导航设计、电子商务等工作变得更为复杂更为繁重。对于网站经营方来说，他们需要更好的自动辅助设计工具，可以根据用户的访问兴趣、访问频度、访问时间动态的调整页面结构，改进服务，开展有针对性的电子商务以更好的满足访问者的需求。解决这种需求的一个有利的工具就是 Web 数据挖掘，即将数据挖掘的思想和方法应用到 Web 上，进行 Web 挖掘，挖掘出有用的信息。

1. Web 数据挖掘概述

Web 挖掘是一项综合技术，涉及 Web、数据挖掘、计算机语言学、信息学等多个领域。Web 挖掘就是从 Web 文档、Web 活动中抽取感兴趣的、潜在的有用模式和隐藏信息。我们从更为一般的角度出发，对 Web 挖掘作如下定义。定义 1 （Web 挖掘） Web 挖掘是指从大量 Web 文档结构和使用的集合 C 中发现隐含的模式 p 。如果将 C 看作输入，p 看作输出，那么 Web挖掘的过程就是从输入到输出的一个映射 ξ：C→p

1.1 与传统的数据挖掘相比较

1.1.1 数据源具有很强的动态性。web 是一个不断变化的、动态更新的系统，web 上的数据信息也是不断更新的。因此，其数据源具有很强的动态性。

1.1.2 挖掘目的的模糊性。web 上有成千上万的用户，而每个用户的背景、使用挖掘的目的和兴趣度都不同，大多数用户对自己的挖掘主题和应用只有一个肤浅的认识和了解，并不能提出一个明确的目标。所以挖掘目的是模糊的、不明确的。

1.1.3 数据类型的多态性。web 上的数据既有数值型数据，也有布尔型数据，还有描述性数据以及 web 特有的数据（如 IP 地址）。新数据类型的出现，必然要对传统的数据挖掘方法进行补充和扩展，才能进行有效的数据挖掘。

1.1.4 数据信息的分布性、多维性。

1.2 Web 数据的特点

Web 技术做为 Internet 飞速发展的产物，对信息在社会中的传播起到了很重要的作用，分布于各 Web 站点上的数据有其自身的特点，主要的可以归纳为以下几点：

1.2.1 数据量巨大。Internet 把分布于世界不同位置的电脑（服务器）连接了起来，每个电脑上都存有丰富的数据，这些数据涉及各种不同的行业和领域，又由于连接于 Internet 的电脑数量非常巨大，所以 Web 站点中的数据量也非常巨大。

1.2.2 异构数据库环境。从数据库研究的角度来看，Web 网站上的所有信息也可以看作是一个比普通数据库更大、更复杂的数据库。每一个 Web 站点都可以看作是一个数据源，由于各站点是相互独立的，之间除了可以互相访问之外并没有任何关系，所以每个站点之间的信息及信息组织方式都是不相同的，这就构成了一个巨大的异构数据库环境。要对这些数据进行分析，必须要解决各站点之间异构数据的集成问题，提供给用户一个统一的视图，才可能从巨大的数据资源中获取有用的信息。

1.2.3 半结构化的数据结构。Web 上的数据与传统数据库中的数据不同之处还在于传统数据库都有一定的模型，可以根据数据模型来对具体的数据进行描述，而 Web 站点中的数据不存在统一的模型，各站点都是独自设计，并且站点中的数据是处于不停变化之中的。虽然 Web 有自身的结构，大体上站点的结构差异并不是特别大，所以可以认为 Web 数据是一种半结构化的数据，这是 Web 数据的另一个重要的特点。

2. Web 数据挖掘相关技术

因为 Web 挖掘应用非常广泛，所以对 Web 挖掘相关技术的研究也很多，针对上述不同类别的 Web 挖掘，有不同的相关技术，下面分别介绍。一般地，Web 挖掘可以分为三类：Web 内容挖掘（Web content mining）、Web 结构挖掘（Web structure mining）、和 Web 使用模式的挖掘（Web usage mining）。

2.1 技术分类

2.1.1 Web 内容挖掘。Web 内容挖掘是从文档内容或其描述中抽取知识的过程。Web 文本文件内容挖掘，基于概念索引的资源发现，以及基于代理的技术都属于这一类。

2.1.2 Web 结构挖掘。Web 结构挖掘是从 WWW 的组织结构和链接关系中推导知识。由于文档之间的互连 WWW 能够提供除文档内容之外的有用信息。利用这些信息，可以对页面进行排序发现重要页面。这方面的代表有 PageRank[] 和 CL EVER ，此外，在多层次 Web 数据仓库（MLDB）中也利用了页面的链接结构。

2.1.3 Web 使用挖掘。Web 使用挖掘的主要目标是从 Web 的访问记录中抽取感兴趣的模式。WWW 中每个服务器保留了访问日志，记录关于用户访问和交互的信息。分析这些数据可以帮助理解用户的行为从而改进站点的结构，或为用户提供个性化的服务。本文研究的重点就在 Web 使用模式挖掘上。

2.2 Web 数据挖掘研究领域及发展

2.2.1 Web 数据挖掘的研究领域类型根据对 Web 数据的感兴趣程度不同，Web 挖掘一般可以分为三类：网络内容挖掘（Web Content mining）、网络结构挖掘（Web structure mining）、网络用法挖掘（Web usage Mining）

2.2.2 网络内容挖掘网络信息内容是由文本、图像、音频、视频、元数据等形式的数据组成的。

2.2.3 网络结构挖掘网络结构挖掘就是挖掘 Web 潜在的链接结构模式。

2.2.4 网络用法挖掘网络内容挖掘和网络结构挖掘的挖掘对象是网上的原始数据，而网络用法挖掘面对的则是在用户和网络交互的过程中抽取出来的第二手数据，包括网络服务器访问记录、代理服务器日志记录、浏览器日志记录、用户简介、注册信息、用户对话或交易信息、用户提问方式等。通过网络用法挖掘，可以了解用户的网络行为数据所具有的意义。

2.3 Web 数据挖掘的四个步骤

2.3.1 查找资源：任务是从目标 Web 文档中得到数据，值得注意的是有时信息资源不仅限于在线 Web 文档，还包括电子邮件、电子文档、新闻组，或者网站的日志数据甚至是通过 Web 形成的交易数据库中的数据。

2.3.2 信息选择和预处理：任务是从取得的 Web 资源中剔除无用信息和将信息进行必要的整理。例如从 Web 文档中自动去除广告连接、去除多余格式标记、自动识别段落或者字段并将数据组织成规整的逻辑形式甚至是关系表。

2.3.3 模式发现：自动进行模式发现。可以在同一个站点内部或

高浓度二硝基重氮酚废水的处理工艺及综合利用

张凯　陆海东
贵州盘江化工厂　551400

[摘　要]工业雷管生产厂在生产起爆药的过程中产生一定量的高浓度二硝基重氮酚废水,处理工艺不单要从环保的角度考虑,而且还要考虑处理工艺的经济实用。湿性二硝基重氮酚在 95℃下加热 240 小时,它的重氮基才完全分解。可利用二硝基重氮酚在水中的热稳定性来考虑二硝基重氮酚废水处理工艺,用锅炉加热的工艺来处理二硝基重氮酚废水,这样处理,不仅工艺简单,而且费用低,并且锅炉产生的蒸汽还可用在其它的工业生产中。

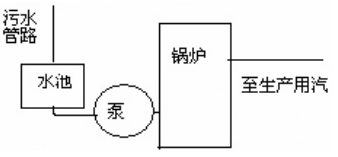
[关键词]高浓度二硝基重氮酚　热稳定性　锅炉加热

1 前言

工业雷管生产厂在生产起爆药(即二硝基重氮酚,分子式 $C_6H_2(NO_2)_2N_2O$)的过程中产生一定量的工艺废水,其中含有较多的二硝基重氮酚。二硝基重氮酚简称 DDNP,干燥状态下的二硝基重氮酚极易爆炸,在常温下微溶于水、挥发性极小。二硝基重氮酚还有一定的毒性,对人主要是中枢神经刺激,中毒症状是身体青紫、头昏目眩、意识不清。如果将废水直接排入工厂的排污沟,将导致工厂周围水源的污染,使人畜引用后中毒。对这种高浓度二硝基重氮酚废水的处理不单从环保的角度考虑,而且还要考虑处理工艺的经济实用。目前国内工业雷管生产厂对二硝基重氮酚废水的处理方法多样,有的还用很原始的自然沉淀和日晒的方法。这种方法虽然经济,但效率太低。二硝基重氮酚能和很多种无机物反应,二硝基重氮酚在比较强的还原剂作用下其中的氮能全部变成氮气放出来,比如三氯化钛和酸。但是利用这种化学的方法——还原反应来处理二硝基重氮酚废水很不经济。我们可以利用二硝基重氮酚在水中的热稳定性来考虑二硝基重氮酚废水处理工艺。

2 二硝基重氮酚废水处理工艺原理及工艺流程

二硝基重氮酚在水中的热稳定性比较好,在 60℃下长时间加热也不见分解现象,在 75℃下加热 960 小时,其损失量仅 0.5%,在 95℃下加热 240 小时,它的重氮基才完全分解。因此,可以用蒸汽锅炉加热的工艺来处理二硝基重氮酚废水,这样处理,不仅工艺简单,而且费用低,并且锅炉产生的蒸汽还可用在其它的工业生产中。其工艺流程如下:



3 工艺简述

二硝基重氮酚废水由管道输送至储存水池,再由水泵连续供给蒸汽锅炉。废水经过连续加热,二硝基重氮酚完全分解,并且锅炉产生的蒸汽可以供给其它生产工艺用汽。二硝基重氮酚废水分解后产生氮气和一些沉淀物,氮气可随水蒸汽一起排出,沉淀物则附着在锅筒壁上。沉淀物物理性能稳定,而且无毒无害,可定期清理后运出,为便于清理沉淀物,最好选择卧式手烧炉。按锅炉蒸发量 4 吨/小时的处理能力,每天处理量达到 96 吨,也可以根据废水的产生量选择相应蒸发量的锅炉。

4 结论

这种利用蒸汽锅炉加热的工艺来处理二硝基重氮酚废水的方法的确是实用可行的。该工艺主要能耗是煤炭,它产生的蒸汽是可以循环再利用的。

参考文献:
1,《工程雷管》国防工业出版社,1977 年
2,《基础化学工程》上海科学技术出版社,1978 年

在多个站点之间进行。

2.3.4 模式分析: 验证、解释上一步骤产生的模式。可以是机器自动完成, 也可以是与分析人员进行交互来完成。

3. Web 数据挖掘的应用

据国外专家预测, 在今后的 5—10 年内, 随着企业数据量的日益积累和中国经济的高速发展, 数据挖掘将在中国形成一个产业。目前国内数据挖掘专业的人才培养体系尚不健全, 人才市场上精通数据挖掘技术、商业智能的专业人才极少, 而另一方面企业、政府机构和科研单位对此类人才的潜在需求量又很大, 此类人才供需量差距较大。网络信息挖掘的应用涉及到电子商务、网站设计和搜索引擎服务等众多方面, 这项技术的应用正变得越来越广泛。

3.1 Web 挖掘在搜索引擎方面的应用

Web 挖掘是目前网络信息检索发展的一个关键, 我们多数的人都是通过搜索相关网页获得信息。通过对网页内容的挖掘, 可以实现对网页的聚类和分类, 实现网络信息的分类浏览与检索。运用 Web 挖掘技术改进关键词加权算法, 提高网络信息的标引准确度, 改善检索效果。参与搜索服务市场的有多家实力企业, 如 Google、雅虎 (Yahoo!) 及微软 (Microsoft) 等巨头企业, 以及若干规模较小但有特定市场区隔或技术者如 dTSearch、Copernic 等。在十几年前, 比尔?盖茨向世人宣誓, “把信息技术带到每个人的指尖”。然而, 信息搜索技术却被盖茨当成了不可能赢利的技术之一, 他的一时大意成就了 Google 公司今天的功勋。Google 站在前人的肩上, 对搜索引擎进行了颠覆传统的修改, 创造出了新的价值, 同时还创造出了一家市值达上千亿美元的公司, 也促使搜索成为互联网的心脏。

3.2 Web 挖掘在电子商务方面的应用

Web 挖掘这方面的应用可以为企业更有效的确认目标市场、改进决策获得更大的竞争优势提供帮助, 从中可得到商家用于特定消费群体或个体进行定向营销的决策信息。电子商务方面的 Web 挖掘功能主要是如下几个方面:首先, 客户分类和客户聚类。其次是找到潜在的客户。在对 Web 的客户访问信息的挖掘中, 利用分类技术可在因特网上找到未来的潜在客户。最后保留客户的驻留时间, 对于客户而言, 传统客户与销售商之间的空间距离在电子商务中已经不复存在, 在网上, 每个销售商对于客户来说都是一样的, 如何尽量使客户在自己的网上驻留更长的时间, 这样对于商家才能有更多客户和更大的利润空间。

3.3 在网站设计中的应用

在网站设计方面中的应用, 主要是通过对网站内容的挖掘, 特别是对文本内容的挖掘, 可以有效地组织网站信息, 如采用自动归类技术实现网站信息的层次性组织; 通过对用户访问日志记录信息的挖掘, 把握用户感兴趣的信息, 从而有助于开展网站信息推送服务以及个人信息的定制服务, 吸引更多的用户。

4.结束语

社会的发展越来越离不开信息的传播与使用,在数据量急剧增长的情况下如何高效地检索出使用者需要的信息更加显得重要,Web 数据挖掘正是因为满足了这方面的需要才能获得如此迅速的发展, Web 挖掘技术也将成为重要的研究课题和方向。

参考文献
[1] 曼丽春, 朱宏, 杨全胜. Web 数据挖掘研究与探讨 [J] .现在电子技术 2005 (8) :3~6
[2] 夏火松. 数据仓库与数据挖掘技术 [M] . 科学出版社,2004.207- 227.
[3] Jiawei Han,Micheline Kamber.DataMining:Concept and Techniques [M] .Morgan Kaufmann Publishers,Inc 2001.272- 312.
[4] 陈文伟,黄金才,赵新昱.数据仓库与数据挖掘技术 [M] .北京:北京大学出版社,2002.1- 14.
[5] 王继成,潘金贵.Web 文本挖掘技术研究 [J] .计算机研究与发展,2000,37 (5) :513- 520.