Web 数据挖掘技术及工具研究

邓 英 李 明
(甘肃工业大学电气工程与信息工程学院,兰州 730050)
E-mail:imtsai@email.com.cn

摘 要 Internet 应用的普及使得数据挖掘技术的重点已经从传统的基于数据库的应用转移到了基于 Web 的应用。文章就 Web 挖掘技术的概念、分类及文本挖掘和用户访问模式挖掘的实现技术做了详细的阐述,并在此基础上介绍了一些实用的 Web 挖掘工具。

关键词 Web 挖掘 文本挖掘 用户模式挖掘

文章编号 1002-8331-(2001)20-0092-03 **文献标识码** A **中图分类号** TP311

Research on Web Mining and Tools

Deng Ying Li Ming
(College of Electrical and Information Engineering,Gansu Univ.Tech,Lanzhou 730050)

Abstract: With the prevalence of Internet application,the emphases on the research of the data mining technology has transferred from DB to the Web.This paper elaborates the conception,classification of the Web mining and technology of text mining and usage mining,then introduces some applied tools of Web mining.

Keywords: Web mining,text mining,Usage mining

1 引言

随着 Internet/Intranet 技术的发展,尤其是 Web 的全球普及,使得 Web 上信息量无比丰富,如何从非格式化数据信息中有效地挖掘出有用的信息是对数据挖掘领域的一个新挑战。

Web 上[1]的数据信息不同于数据库。数据库有规范的结构,如关系数据库的二维表结构。它有统一的格式,其中的数据为完全结构化的数据。Web 上的信息则不然,主要是些大量的、异质的 Web 信息资源,文档结构性差,其数据多为半结构化或非结构化。由于半结构化和非结构化的信息不能清楚地用数据模型来表示,因此在 Web 上的数据挖掘需要用到很多不同于单个数据仓库挖掘的技术。文章将对 Web 挖掘技术做系统性的研究和探讨,并在此基础上介绍一些用于 Web 挖掘的工具。

2 Web 挖掘概念

Web 挖掘是利用数据挖掘技术从 Web 文档及 Web 服务中自动发现并提取人们感兴趣的信息。它是一项综合技术,涉及到 Internet 技术、人工智能、计算机语言学、信息学、统计学等多个领域。通常 Web 挖掘过程可以分为以下几个处理阶段:资源发现、数据抽取及数据预处理阶段、数据汇总及模式识别阶段、分析验证阶段。

不同研究者从自身的领域出发,对 Web 挖掘的含义有着不同的理解,因此项目开发也各有其侧重点。通常人们往往将 Web 挖掘与 Web 上的信息检索或信息抽取等同起来,其实,它们之间是有区别的,主要体现在:(1)网络信息检索系统只能处理以关键词形式表示的简单目标,无法处理用户给出的样本形式的复杂模糊目标,而挖掘系统则能够从文本中提取出目标信息的特征,然后根据目标特征在网络中进行有目的的搜寻,将搜寻到的文档提交给用户。(2)信息检索目的是针对某一特定领域进行信息或文档的收集,可以看作是用于 Web 挖掘中文档分类的一种情况。(3)不是所有的信息检索都要用到数据挖掘技术,因此信息检索通常不能发现隐藏在数据后面的联系,而 Web 挖掘的目的就是将大量看似无关的数据关联起来发现其中的规则和知识以供决策支持。尽管 Web 挖掘不同于信息检索,但它们在实现技术上却有很多相似之处,所以 Web 挖掘技术可以借鉴信息搜索技术。

Web 挖掘可分为三类:内容挖掘、结构挖掘、用户访问模式挖掘,而 Web 信息挖掘和用户访问模式挖掘是 Web 挖掘的两个主要方面。文章就这两个主要方面进行论述。

2.1 Web 内容挖掘

Web 的内容挖掘可以说是将数据挖掘技术在网络信息处理中的应用,不同于传统的数据挖掘技术,Web 挖掘主要是针对各种非结构化的数据,如文本数据、音频数据、视频数据、图形图象数据等多种数据相融合的多媒体数据挖掘。又可将其分为基于文本的挖掘和基于多媒体的挖掘两种。

基于文本的 Web 挖掘方法有数据库方法,建立 Web 数据仓库方法和新近的基于软件 Agent 的分类器方法、基于概念的文本信息挖掘法。Web 多媒体的信息挖掘通常采用的方法为关联规则法和特征提取法。

数据库方法和数据仓库都是采用数据抽取和转换的方法将非结构化的 Web 信息转化或映射为结构化的数据结构,然

基金项目:甘肃省自然科学基金项目资助(编号 ZS001-A22-021-G)
作者简介:邓英,硕士研究生,研究方向:数据仓库与决策支持系统,数据挖掘,网络数据库。

后就可以采用数据库挖掘技术进行信息挖掘。

2.2 用户模式挖掘

用户使用 Web 获取信息的过程中需要不停地从一个 Web 站点通过超文本链接跳转到另一个站点,这种过程存在一定的普遍性,发现此规律即是 Web 用户访问模式发现。这是一种完全不同于上述所讲的资源发现的任务。它是对现代电子商务战略的一个重要支持。面向 Web 用户访问模式的挖掘是关于用户行为及潜在顾客信息的发现,包括三种模式,即数据预处理、模式发现及模式分析。在此,数据挖掘的主要任务是从数据中发现模式。通常实现方法是对 Sever Logs、Error Logs 和 Cookie Logs 等日志文件的分析挖掘出用户访问行为、频度和内容等信息,从而找出一定的模式和规则。

理解 Web 上的用户访问模式有如下好处:合理建造网站及合理设计服务器,如辅助改进分布式网络系统的设计性能,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;帮助改善市场营销决策,如把广告放在适当的 Web 页上或更好地理解客户的兴趣,这样的知识将有助于商家制定促销策略。

3 Web 挖掘技术研究

Web 挖掘从数据挖掘发展而来,数据挖掘方法通常可分为两类,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、遗传算法等。

3.1 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。Web 文档多为 HTML、XML 等自然语言,因此可以利用 Web 文档中的标记,如〈Title〉,〈Heading〉等额外信息,利用这些信息来提高 Web 文本挖掘的性能。Web 文本挖掘可以对 Web 上大量文档集合的内容进行总结、分类、聚类、关联分析等。

文本总结。其目的是对文本信息进行浓缩,给出它的紧凑描述。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。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。文本总结在有些场合十分有用,例如,搜索引擎在向用户返回查询结果时,通常需要给出文档的摘要。

文本分类。分类的概念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(Classifier)。分类器一般分为训练和分类两个阶段。分类往往表现为一棵分类树,根据数据的值从树根开始搜索,沿着数据满足的分支往上走,走到树叶就能确定类别。

分类器的构造方法有统计方法、机器学习方法、神经网络方法等等。统计方法包括贝叶斯法和非参数法(近邻学习或基于事例的学习),对应的知识表示则为判别函数和原型事例。机器学习方法包括决策树法和规则归纳法,前者对应的表示为决策树或判别树,后者则一般为产生式规则。神经网络方法主要是 BP 算法,它的模型表示是前向反馈神经网络模型(由代表神经元的节点和代表联接权值的边组成的一种体系结构),BP 算法本质上是一种非线性判别函数。

文本聚类。文本聚类是一种典型的无教师的机器学习问题。目前的文本聚类方法大致可以分为层次凝聚法和平面划分法两种类型。聚类是把一组个体按照相似性归成若干类别,即“物以类聚”。它的目的是使得属于同一类别的个体之间的距离尽可能的小,而不同类别上的个体间的距离尽可能的大。

关联规则。关联规则模式属于描述型模式,发现关联规则的算法属于无监督学习的方法。关联规则的定义为:若 X、Y 为项目集,且 $Y \mid X=\Phi$,蕴涵式 $X \Rightarrow Y$ 称为关联规则,X、Y 分别称为关联规则 $X \Rightarrow Y$ 的前提和结果。项目集(XYY)的支持率称为关联规则 $X \Rightarrow Y$ 的支持率,定义为:

$$\text{support}(X \Rightarrow Y)=\text{support}(XYY)$$

关联规则 $X \Rightarrow Y$ 的置信度定义为:$\text{confidence}(X \Rightarrow Y)=\frac{\text{sup prot}(XYY)}{\text{sup port}(X)} \times 100\%$

支持率和置信度是描述关联规则的两个重要概念,前者用于衡量关联规则在整个数据集中的统计重要性,后者用于衡量关联规则的可信程度。通常用户只对支持率和置信度均高的关联规则感兴趣,也只有支持率和置信度均高的关联规则才是有用的关联规则。发现关联规则通常要经过以下三个步骤:

(1)连接数据,作数据准备;

(2)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;

(3)可视化显示、理解、评估关联规则。

文本聚类与分类的不同之处在于,聚类没有预先定义好的主题类别,它的目标是将文档集合分成若干个簇,要求同一簇内文档内容的相似度尽可能的大,而不同簇间的相似度尽可能的小。

由于 Web 的开放性、动态性与异构性等固有特点,要从这些分散的、异构的、没有统一管理的海量数据中快速、准确地获取信息也成为 Web 挖掘所要解决的一个难点,也使得用于 Web 的挖掘技术不能照搬用于数据库的挖掘技术。因此,开发新的 Web 挖掘技术,以及对 Web 文档进行预处理以得到关于文档的特征表示,便成为 Web 挖掘研究的重点。

3.2 用户模式挖掘实现技术

用户访问模式又可称为用户导航信息。在 Web 的用户访问模式的挖掘中,描述用户访问模式的数据包括:IP 地址、参考页面、访问日期和时间、用户的 Web 站点及配置信息。这些数据可以来自于服务器端、客户端、代理服务器端或者是公司的数据库。

常用的有两种方法发现用户导航信息。一种方法是通过对日志文件进行分析,又有两种方式,其一访问前先进行预处理,即将日志数据映射为关系表并采用相应的数据挖掘技术,如关联规则或聚类技术来访问日志数据,其二是对日志数据进行直接访问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。

从研究目标的角度看,已有的基于 Web 服务器日志数据的研究大致可以分为 3 类:(1)以分析系统性能为目标;(2)以改进系统设计为目标;(3)以理解用户意图为目标。由于各目标针对的功能不同,采取的主要技术也不同。

用户导航信息的挖掘通常要经过下面三个步骤:(1)数据预处理阶段。这是用户导航信息挖掘最关键的阶段,数据预处理包括:关于用户导航信息的预处理、关于内容预处理和结构的预处理;(2)模式识别阶段。该阶段采用的方法包括:统计法、机器学习和模式识别等方法。实现算法可以是:统计分析、聚

类、分类、关联规则、序列模式识别等；(3)模式分析阶段。该阶段的任务是从上一阶段收集的数据集中过滤掉不感兴趣和无关联的数据及模式。具体的实现方法要依具体采用的 Web 挖掘技术而定，通常采用的方法有两种：一种采用 SQL 查询语句进行分析；另外一种将数据导入多维数据立方体中，而后利用 OLAP 工具进行分析并提供可视化的结果输出。

对挖掘用户导航信息的研究早期多采用的是统计的方法，当用户通过浏览器对 Web 站点进行访问时，建立统计模型对用户访问模式进行多种简单的统计，如频繁访问页、单位时间访问数、访问数据量随时间分布图等。早期使用的方法为以广度优先算法为主的统计模型，还有一种启发式的 HPG(hypertext probabilistic grammar)模型[4]用于用户导航行为的发现，它也是一种基于统计的方法，由于 HPG 模型与 k 阶马尔可夫模型相当，所以近来也有人提出用马尔可夫模型挖掘用户导航信息。

4 Web 挖掘工具的介绍

在数据挖掘技术日益发展的同时，许多数据挖掘的商业软件工具也逐渐问世。评价一个数据挖掘工具，需要从以下几个方面来考虑：(1)可产生的模式种类的多少；(2)解决复杂问题的能力；(3)易操作性；(4)数据存取能力；(5)与其他产品的接口。

通用的数据挖掘工具有，IBM 公司 Almaden 研究中心开发的 QUEST 系统，SGI 公司开发的 MineSet 系统，加拿大 Simon Fraser 大学开发的 DBMiner 系统。处理特定领域的数据挖掘工具有 IBM 公司的 Advanced Scout 系统针对 NBA 的数据，帮助教练优化战术组合；加州理工学院喷气推进实验室与天文科学家合作开发的 SKICAT 系统，帮助天文学家发现遥远的类星体；芬兰赫尔辛基大学计算机科学系开发的 TASA，帮助预测网络通信中的警报。

上述几种挖掘工具对象可以说主要都是针对结构化的数据进行分析处理，下面主要介绍几种适用于 Web 挖掘的工具。

4.1 文本信息挖掘工具

通常文本挖掘工具主要完成两方面的工作：信息检索和对文本的分析。文本挖掘工具的主要设计目标是使用户用于理解文档内容或用于收集相关文档所花费的时间最少。IBM 公司推出的 Web 文本挖掘工具 Intelligent Miner for Text，它是 IBM 开发的 Intelligent Miner 家族的一成员，它主要包括三部分：高级搜索引擎 TextMiner，其最大特点是具有在线更新的能力，即它在执行索引任务的同时无须将搜索进程挂起，可获得较高的效率；Web 访问工具包括一个优化的搜索引擎 NetQuestion 和 Web Crawler，Web Crawler 是一个可以在一个或多个 Web 站点启动的自动机，它可以监视 Web 页的活动并可以变更检索使之更优化；文本分析工具，这部分完成的才是对文本信息的挖掘，这部分工具可以独立使用，但将它与文本搜索工具结合使用将能发挥更强大的作用。该软件主要是由信息提取器工具组成，该工具提供了高效的文本信息挖掘，可以实现全文搜索、文本分析、Web 文档查询和检索。

4.2 用户访问模式挖掘工具

由 Stephen Turner 博士编制的免费个人软件 Analog，是一个用来分析服务器日志文件的工具，它适用于 Windows 及 UNIX 等操作系统中，由于它的使用较简单，可以直接在服务器上运行，也可以将日志文件下载到客户端，在客户端运行。比较适用于个人和小规模分析应用，是一个实用性很强的日志文件分析工具。从 http://www.statlab.cam.ac.uk/上可免费获得该软件。

用户导航行为挖掘工具 WUM (Web Utilization Miner)是一种序列挖掘器。它主要用来分析用户导航行为的发现，它适用于从任何类型的日志文件中发现用户导航信息。WUM 是一个对日志文件进行集成处理、查询及分析的工具，它的核心是 MINT 处理器，主要是对从 Web 日志文件中提取的集成信息进行分析，从而发现导航模式。MINT 是用于用户和挖掘器接口的语言，这种语言为用户提供了更为强大、灵活和全面的功能，它可以根据用户输入的语法标准进行以用户为前提的分析工具。正是因为 WUM 能提供较强大和灵活的功能，所以对用户也提出了较高的要求。要求用户掌握 MINT 语言，并具有能对挖掘结果进行分析处理所具备的知识。MINT 语言语法是一个包含了 SQL 查询语句中变量和通配符的模板，它与 SQL 查询语言有类似的语法结构，对用户而言比较容易掌握和使用。可从网上免费获得软件 WUM5.0 的演示版本，其网址为：http://wum.wiwi.hu-berlin.de/。

5 总结

文章就 Web 挖掘分类及技术做了概括性的介绍，同时介绍了一些很实用的工具。Web 挖掘还有待进一步的研究，尤其是近来对 Web 内容挖掘方面集中在信息集成，如建立基于 Web 的知识库或基于 Web 的数据仓库的研究上，但这种方法同样存在很多问题，首先是如何从 Web 中提取有用的信息来建造 Web 数据仓库，其次所要提取信息的 Web 站点范围的确定依据如何选取，Web 上的数据结构不同于数据库上的数据结构。但建立一个基于 Web 数据仓库的数据挖掘系统仍是一种值得研究的方法。文献[5]介绍了 Web 数据仓库建立和挖掘的方法，用户访问模式的挖掘近来主要集中在对数据预处理技术[6]及算法的研究方面。(收稿日期：2001 年 7 月)

参考文献

1.王继成，潘金贵等.Web 文本挖掘技术研究[J].计算机研究与发展，2000，37(5)：513-520

2.Stanley Loh,Landro Krug Wives.Concept-Based Knowledge Discovery in Texts Extracted from the Web[J].SIGKDD Explorations，2000，7：29-39

3.David Heckerman.Bayesian Networks for Data Mining[J].Data Mining and Knowledge Discovery，1997，1：79-119

4.Jose Borges,Mark Levene.A Fine Grained Heuristic to Capture Web Navigation Patterns[J].ACM SIGKDD，2000，2(7)：1-40

5.Sanjay Madria，S Bhowmick.Research Issues in Web Data Mining[C].Research Issues in Web Data Mining.9th Intl.Database Conf，Hong Kong，1999.7：13-27

6.刘明吉，王秀峰等.数据挖掘中的数据预处理[J].计算机科学，2000，27(4)：54-57

阅读此文的还阅读了：

1. 国产化可投捞气举采油工具的研制
2. 书斋说禅(之一)
3. 大数据难题：超越技术
4. P3e^TM技术揭示了一个新领域
5. 项目启动阶段工作实践:一项调查研究结果(上)
6. 基于web的数据挖掘技术研究
7. Trawling算法在Web结构挖掘中的应用
8. 定向井分层注水工艺研究与应用
9. 斯大林關於基礎與上層建築學說對財政學研究的啓示
10. 教学评价的理念和技术——普通高中生物课程标准中“评价建议”的介绍
11. 音乐2.0
12. 软件工程的工具、技术及方法
13. web数据挖掘技术的研究
14. Web数据挖掘研究
15. 车辆维修技术的发展新方向分析
16. 基于Web的数据挖掘技术的研究

17. 项目启动阶段工作实践:一项调查研究结果(上)

18. 袋式堵漏工具研制与试验

19. 国外小井眼钻井技术的发展及启示

20. 数据挖掘在提高web用户网络访问速度上的研究

21. 客服电话成为市调工具

22. Web数据挖掘技术的研究

23. 数据挖掘在企业中的应用

24. 数据挖掘技术及工具研究

25. 人类走出非洲并未携带先进工具

26. 数据挖掘技术研究及应用

27. 货币只是技术工具——化解危机还须结构重建

28. 软件工程的工具、技术及方法(二)

29. 2014年1月技术雷达：工具、语言和框架

30. Wiki——建构知识王国助力教育发展

31. 除垢工具和工艺优化研究与应用

32. Web数据挖掘技术研究

33. 论人文技术哲学视野下的教育技术观

34. 基于Web日志的数据挖掘技术在Web机器人识别中的研究

35. Web上的数据挖掘技术和工具设计

36. 井下作业工具的创新与发展研究与分析

37. 基于Web数据挖掘技术研究

38. WEB数据挖掘技术研究

39. Web数据挖掘技术研究综述

40. 谈谈数据仓库及数据挖掘工具

41. 面向Web数据挖掘技术的研究

42. 基于Web的新闻发布管理系统的研究与实现

43. 数据挖掘在企业中的应用

44. 基于Web的数据挖掘技术研究及其应用

45. 马提亚·曼恩:质谱分析与蛋白质组

46. 井下实用新工具发展现状及技术研究

47. 第三部分新一代工具与技术

48. 数据挖掘工具的评判

49. 利用矿业软件Micromine研究与实践

50. 基于web的数据挖掘系统的研究与设计