

电子商务中 Web 挖掘技术的应用探讨

但 微¹, 才书训²

(1. 东北大学 软件学院, 辽宁 沈阳 110004;

2. 东北大学 秦皇岛分校, 河北 秦皇岛 066004)

摘 要:随着电子商务技术的深入发展,商家在与用户交互活动中的信息也迅速膨胀,网上的海量信息为 Web 挖掘提供了一个广阔的应用领域,使用 Web 挖掘技术能够发现电子商务过程中的潜在模式。文中针对几种不同特点的电子商务 Web 挖掘系统,包括智能搜索引擎系统、网站用户访问分析系统、个性化推荐系统等都进行了探讨。

关键词:Web 挖掘;智能搜索引擎;网站用户访问分析;个性化推荐

中图分类号:TP391.3;F713.36

文献标识码:A

文章编号:1005-3751(2006)01-0207-03

Using Web Mining in Electronic Commerce

DAN Wei¹, CAI Shu-xun²

(1. Software College, Northeastern University, Shenyang 110004, China;

2. Qinhuangdao Branch, Northeastern University, Qinhuangdao 066004, China)

Abstract: With the great development of electronic commerce, activity and knowledge exchanging between enterprises and customers have increased. The World Wide Web is a fertile area for Web mining. This article involved the different characteristics of Web mining system, including intelligent search engine system, Web usage analysis system and personalized recommendation system.

Key words: Web mining; intelligent search engine; Web usage analysis; personalized recommendation

0 引 言

Web 挖掘是针对包括 Web 页面内容、页面之间的结构、用户访问信息、电子商务信息等在内的各种 Web 数据源,在一定基础上应用数据挖掘的方法以发现有用的隐含的知识的过程。Web 挖掘与传统的数据挖掘相比有其自身的特点,Web 本身是半结构化或无结构的数据,缺乏机器可理解的语义,Web 挖掘的对象是大量、异质、分布的 Web 文档,对 Web 服务器上的日志、用户信息等数据所开展的挖掘工作也属于 Web 数据挖掘的范畴。目前已经出现了许多电子商务领域的 Web 挖掘系统应用。文中首先介绍 Web 挖掘技术所能实现的功能,然后在此基础上介绍在电子商务领域广泛应用的智能搜索引擎、网站用户访问分析和个性化推荐中 Web 挖掘技术应用的情况。

1 Web 挖掘技术的功能介绍

电子商务 Web 信息的多样性决定了挖掘任务的多样性。按照 Web 处理对象的不同,一般将 Web 挖掘分为 3 类:Web 内容挖掘、Web 结构挖掘和 Web 使用记录挖掘(如图 1 所示)。针对这 3 种不同的处理对象,能够挖掘出

许多有用的信息。

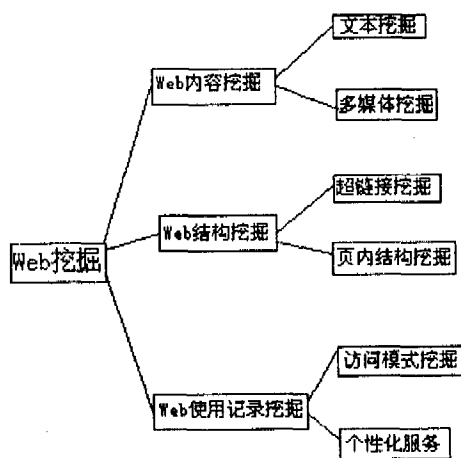


图 1 Web 挖掘的分类

1.1 Web 内容挖掘

Web 内容挖掘是指从文档的内容中提取知识。Web 内容挖掘又分为文本挖掘和多媒体挖掘。目前多媒体数据的挖掘研究还处于探索阶段,Web 文本挖掘已经有了比较实用的功能。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,以及利用 Web 文档进行趋势预测等。Web 文档中的标记,例如 < Title> 和 < Heading> 等蕴含了额外的信息,可以利用这些信息来加强 Web 文本挖掘的作用。

收稿日期:2005-04-10

作者简介:但 微(1981—),男,江西都昌人,硕士研究生,研究方向为电子商务及 Web 数据挖掘。

1.2 Web 结构挖掘

Web 结构挖掘是从 Web 的组织结构和链接关系中推知知识。它不仅仅局限于文档之间的超链接结构,还包括文档内部的结构、文档中的 URL 目录路径的结构等。Web 结构挖掘能够利用网页间的超链接信息对搜索引擎的检索结果进行相关度排序,寻找个人主页和相似网页,提高 Web 搜索蜘蛛在网上的爬行效率,沿着超链接优先爬行。Web 结构挖掘还可以用于对 Web 页进行分类、预测用户的 Web 链接使用及 Web 链接属性的可视化、对各个商业搜索引擎索引的页数量进行统计分析等。

1.3 Web 使用记录挖掘

Web 使用记录挖掘是指从 Web 的使用记录中提取感兴趣的模式。目前 Web 使用记录挖掘方面的研究较多,WWW 中的每个服务器都保留了访问日志,记录了关于用户访问和交互的信息,可以通过分析和研究 Web 日志记录中的规律,来识别电子商务的潜在客户^[1];可以用基于扩展有向树模型来识别用户浏览序列模式,从而进行 Web 日志挖掘;可以根据用户访问的 Web 记录挖掘用户的兴趣关联规则,存放在兴趣关联知识库中,作为对用户行为进行预测的依据,从而为用户预取一些 Web 页面,加快用户获取页面的速度。分析这些数据还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。

2 智能搜索引擎

搜索引擎已成为人们上网浏览时的重要工具,用户通过搜索引擎在浩瀚的网站页面的海洋中迅速地找到自己所需的信息。这种市场的需求刺激着搜索引擎的技术不断地走向完善,将新的 Web 挖掘技术应用到传统的搜索引擎中去,引起了搜索引擎业界的一场革命,以下便是 Web 挖掘在智能门户搜索引擎中的一些应用。

2.1 文档自动分类

目录式搜索引擎和机器人搜索引擎各有利弊,应用 Web 挖掘技术,可以有效地改善其性能。搜索引擎通过向 Internet 发送搜索蜘蛛机器人程序自动地从所爬行过的网页上抽取检索到的信息,然后连同该网页的 URL 地址一起存入搜索引擎的索引数据库中。通过 Web 挖掘可以对索引数据库中的信息进行整理,对文档进行自动分类,从而提高了用户的检索速度和检索的精确度。

2.2 自动文摘的形成

搜索引擎在向用户返回检索结果时,通常要给出每个文档的一个简单的摘要。目前,大部分搜索引擎是机械地截取文档的前几句。利用 Web 文本挖掘中的文本总结技术,可以从 Web 页中提炼出重要信息形成文档摘要,使用户能快速、方便、准确地了解检索信息。

2.3 检索结果的联机聚类

用户使用搜索引擎时会得到由大量的返回信息组成的线性表,其中很大一部分是与用户的查询请求不相关

的。通过对检索结果的文档集合进行聚类,可以使得与用户检索结果相关的文档聚类得比较靠近,从而远离那些不相关的文档。将处理以后的信息以超链接结构组织的层次方式可视化地提供给用户,由用户选择他所感兴趣的那一簇,大大缩小了所需浏览的页面数量。

2.4 查询结果的相关度排序

Web 链接结构中包含了许多有用的信息,当一个网页的作者建立了一个指向其他网页的链接时,就可以看作是作者对其他网页的引用。因此,指向一个文档的链接体现了该文档的被引用情况。如果大量的链接都指向了同一个网页,就认为它是一个权威页。按照这种思路就可以使用 Web 挖掘技术对搜索结果进行相关度排序^[2]。利用链接文本对被引用的页面进行索引,链接文本经常给出比页面本身更加精确的页面描述,而且借助于链接文本的存在,可以检索出一些非文本文档,如图像、程序、数据库等。另外,可利用文档其他的可见描述信息,例如认为字号大或黑体字比其它单词具有更大的权重。

3 网站用户访问分析

3.1 系统对网站用户访问的分析项目

3.1.1 网站的概要统计

网站的概要统计包括分析覆盖的时间、总的页面数/访问数/会话数/唯一访问者等,以及平均访问、最高访问、上周访问、昨日访问等结果集。

3.1.2 访问分析

访问分析包括最多/最少被访问的页面、最多访问路径、最多访问的新闻、最高访问的时间、最多访问的动态页面。

3.1.3 客户信息分析

客户信息分析包括访问者的来源省份统计、访问者使用的浏览器/操作系统分析、访问来自的页面或者网站、来自的 IP,以及访问使用的搜索引擎。

3.1.4 访问者活动周期行为分析

访问者活动周期行为分析包括一周内 7 天的访问走势、一天内 24 小时的访问走势、每周的最多的访问日、每天的最多访问时段等。

3.1.5 主要访问错误分析

主要访问错误分析包括服务端错误、页面找不到错误等。

3.1.6 网站栏目分析

网站栏目分析包括定制的频道和栏目设定,统计出各个栏目的访问情况,并进行分析。

3.1.7 商务网站扩展分析

商务网站扩展分析是专门针对专题/多媒体文件/下载等内容的访问分析。

3.2 商用的网站用户访问分析系统

已经有很多公司开发出了商用的网站用户访问分析系统,以下是几个著名的 Web 数据挖掘方向的公司及其

产品的介绍。

3.2.1 WebTrends 公司

它的重要产品是 CommerceTrends 3.0,它能够让电子商务网站更好地理解其网站访问者的行为,帮助网站采取一些行动来将这些访问者变为顾客,将一次性的顾客变为长期的忠实顾客。CommerceTrends 提供了“browser-based”的方法,使得不同的部门(从市场部门到分析家)都能得到个性化的报表。

CommerceTrends 主要由如下 3 部分组成。

(1)报表生成服务器:它提供相关的 Web 流量信息。这些报表能够自动生成,也可以根据要求实时地生成,能够提供天、星期、月等的总结性的报表,实现了动态与静态的结合。

(2)Campain Analyzer:网站的浏览者或者是看一眼就走,或者表现出很强烈的兴趣,网站的经营管理者可以根据这些差别分析原因,从而制订正确的市场战略。

(3)Webhouse Builder:它能够提供可利用的数据,根据这些数据来产生访问者的行为模式。而且它将其其他技术例如 CRM, ERP, personalization solution 融合到这里,因此对访问者和他们的行为有一个比较完全的理解。

3.2.2 Accrue 公司

该公司的产品 Accrue Insight 是一个综合性的 Web 分析工具,它能够对网站的运行状况有个深入、细致和准确的分析,通过分析顾客的行为模式,帮助网站采取措施来提高顾客对于网站的忠诚度,从而建立长期的顾客关系。Accrue Insight 利用了多种 Web 数据收集方法,能够收集到 Web server 日志里所得不到的信息,例如按下“停止”键、下载的时间等一些对于网站分析有用的信息。但是对于加密的 Session 或者遇到它不适用的部分则用到另外的方法。根据原始数据,Accrue Insight 运用了一种叫做“Server Collector”的分析方法,它支持镜像服务器和负载均衡、路由器和一些其他网络结构设备,能够将一些加密的地址转化为可分析的形式。

3.2.3 Net Perception 公司

这个公司主要的产品是 Net perceptions,它采用了一个叫做“实时建议”的技术,即让它的产品对象(主要是网站)能够根据用户以往的浏览行为(包括这次的 Click-throughs 和以前的购买记录),在其他用户(称为 Community)中找出与他有相类似浏览行为的,根据这些用户的浏览行为来预测该用户以后的浏览行为,从而为用户提供个性化的浏览建议,例如 Cross-sell 和 Up-sell。这种技术利用了网站用户的浏览行为有相似的一面,因此其预言有很高的准确性。并且它是实时运行的,随着浏览量的增加,它会变得越来越“聪明”。

4 个性化推荐

通过 Web 挖掘得到用户的兴趣和爱好,并以此进行个性化推荐,是目前电子商务系统建设所采用的一种重要

手段。分析用户使用记录数据可以帮助系统管理者理解用户的行为,得到用户群体普遍的访问行为模式和用户个体的访问模式,从而根据这种模式为用户定制合适的推荐页面。

在个性化推荐系统中,访问页面关联规则和访问模式聚类分析是两种常用的技术^[3]。基于用户访问页面的关联规则挖掘应用于推荐,是采用精确的访问模式匹配,推荐准确率高;基于用户访问模式的聚类规则应用于推荐,是采用模式相似性原则,推荐覆盖率高。

个性化推荐系统根据其所采用的推荐实现技术又可以分为两种:基于规则的系统和信息过滤系统。信息过滤系统又可分为基于内容过滤的系统和协作过滤系统^[4]。

基于规则的系统:它们允许系统管理员根据用户的静态特征和动态属性来制定规则,一个规则本质上是一个 If-Then 语句,规则决定了在不同的情况下如何提供不同的服务。基于规则的系统其优点是简单、直接,缺点是规则质量很难保证,而且不能动态更新,但是,随着规则的数量增多,系统将变得越来越难以管理。IBM WebSphere (www.ibm.com/websphere)、ILOG (www.ilog.com) 和 BroadVision (www.broadvision.com) 等都是基于规则的推荐系统。

基于内容过滤的系统:它们利用资源与用户兴趣的相似性来过滤信息。比如在电子商务网站中,如果一个用户多次购买某个著名作者出版的书籍,那么当这个作者出版的新书到货时,系统就能够推荐这本新书给这个用户。基于内容过滤的系统其优点是简单、有效,缺点是难以区分资源内容的品质和风格,而且不能为用户发现新的感兴趣的资源,只能发现和用户已有兴趣相似的资源。WebPersonalizer^[5]和 IfWeb^[6]都是典型的基于内容过滤的推荐系统。

协作过滤系统:它们利用用户之间的相似性来过滤信息。基于协作过滤系统的优点是能为用户发现新的感兴趣的信息,缺点是在系统使用初期,由于系统资源还未获得足够多的评价,系统很难利用这些评价来发现相似的用户。SiteSeer (www.rocky.bms.umist.ac.uk/SiteSeer/) 以及 LikeMinds (www.macromedia.com) 和 GroupLens (www.cs.unm.edu/research/GroupLens/index.html) 等都是协作过滤系统。

还有一些个性化推荐系统同时采用了基于内容过滤和协作过滤这两种技术,结合这两种过滤技术可以克服各自的一些缺点,可以利用用户浏览过的资源内容预期用户对其他资源的评价,这样可以增加资源评价的密度,利用这些评价再进行协作过滤,从而提高协作过滤的性能。

5 总结

Web 挖掘技术已经能够对 Web 数据进行内容挖掘、结构挖掘和用户使用记录挖掘,从而调整页面结构,改进

(下转第 216 页)

收到各检测代理的心跳信息后,对比不同检测代理处理数据的总量,如果有个别代理点处理的数据量与全部代理点处理数据量的平均值差距超出了限定的差距阈值,则监控中心会在平均值两侧各找一个差距较大的代理点对其进行平衡(即最佳适应)。

平衡时以高负载点向低负载点让出部分负载的形式实现。平衡的最终目标是两个检测代理点的负载都接近或等于系统检测代理负载平均值。监控中心首先向两个平衡点各发送一个带有标志位的平衡点数据,标志位标示本代理点为高负载代理或低负载代理点,数据中包含了参与平衡的另一代理点的 IP 和整个系统各检测代理的负载平均值(即中央控制策略)。此后监控中心不再参与平衡活动,整个平衡活动以高负载的一方为中心进行(即自主控制策略)。高负载代理点从自身的负载中取出一部分负载,将这些负载的信息和在本检测点上存储的负载的历史统计信息传送给低负载点,由低负载检测代理点接收并开始检测工作。低负载点在启动对这部分的检测后发送完成信号给高负载检测点和检测中心,收到完成信号后,高负载点正式中止对这部分的检测,检测中心则修改分配表中的信息。

在平衡过程中,如果高负载点让出大数据量的被检测负载则需要传输的负载信息一般较大,而如果让出数据量较小的负载则可能会造成两检测代理点的负载点数量相差较大。因此应该从系统负载平均值附近选择适量的负载点。同时为了完成以上的工作各检测代理点应该维护一个记录被检测目标与其数据量的值的对应表(见表 1)。

表 1 被检测点数据量记录表

| | |
|---------|-------|
| 被检测目标 1 | 数据量 |
| 被检测目标 2 | 数据量 |
| 被检测目标 3 | 数据量 |
| | |

(上接第 209 页)

服务,给客户更加个性化的界面,开展有针对性的电子商务,以更好地满足访问者的需求。文中主要介绍了智能搜索引擎系统、网站用户访问分析系统和个性化推荐系统 3 种不同特点的电子商务 Web 挖掘系统。但是,Web 挖掘系统的应用还处于比较初级的阶段,如何让 Web 挖掘系统有更高的智能,还有很多问题需要解决。

参考文献:

[1] 王玉珍. Web 使用模式挖掘在电子商务中的应用[J]. 计算机应用研究, 2003, 10: 155-157.
[2] 李岩, 陈新中, 杨炳儒. 基于 Web 挖掘的智能门户搜索引擎的研究[J]. 计算机工程与应用, 2002, 38(4): 34-36.
[3] 鲍玉斌, 王大玲, 于戈. 关联规则和聚类分析在个性化

(5)如果在一个心跳周期后检测中心未能收到检测代理点发来的心跳信息,则认为检测代理点未能正常工作,此时检测中心根据分配表中的数据将故障代理的负载分配到其他的有效检测点上去,同时修改分配表并且向管理员报警,从而保证了系统整体的有效性和安全性。

(6)为保证系统内部的数据传输高效性,将系统各部分传输的数据设置为统一格式^[8](如图 3 所示),以数据类型标志位来区分不同的数据。

| | | |
|---------|---------|----|
| 底层数据包首部 | 数据类型标志位 | 数据 |
|---------|---------|----|

图 3 系统数据格式

6 总结

通过对以上策略进行有针对性的实现并在中型网络上进行测试实验,在试验过程中合理选取各种阈值,发现以上的设计对分布式入侵检测系统的整体性能有明显提高,同时也提高了系统整体的安全性。

参考文献:

[1] 张世永. 网络信息安全技术[M]. 北京: 科学出版社, 2003.
[2] 张铭来, 金成彪, 赵文耘. 网络型入侵检测系统存在的漏洞及其对策研究[J]. 计算机工程, 2002, 28(1): 172-174.
[3] 史美林, 钱俊, 董永乐. 入侵检测技术与其发展趋势[R]. 北京: 清华大学计算机系 CSCW 实验室, 2002.
[4] 纪祥敏, 连一峰, 许晓利. 入侵检测技术的研究与进展[J]. 计算机仿真, 2004, 21(11): 129-132.
[5] 王晓程, 刘恩德. 攻击分类与分布式网络入侵检测系统[J]. 计算机研究与发展, 2001(6): 17-20.
[6] 陈华平. 分布式动态负载平衡调度的一个通用模型[J]. 软件学报, 1998, 9(1): 25-29.
[7] 唐正军. 网络入侵检测系统的设计与实现[M]. 北京: 电子工业出版社, 2002.
[8] Comer D E. Internetworking With TCP/IP[M]. 北京: 电子工业出版社, 2003.

推荐中的应用[J]. 东北大学学报(自然科学版), 2003, 24(12): 1149-1152.
[4] 曾春, 邢春晓, 周立柱. 个性化服务技术综述[J]. 软件学报, 2002, 13(10): 1952-1961.
[5] Mobsher B. WebPersonalizer: a server side recommender system based on web usage mining[EB/OL]. <http://www.cs.depaul.edu/research/technical.asp>, 2001.
[6] Asnicar F, Tasso C. ifWeb: a prototype of user model based intelligent agent for documentation filtering and navigation in the World Wide Web[A]. In: Tasso C, Jameson A, Paris C L. Proceedings of the UM 1997 Workshop on Adaptive Systems and User Modeling on the World Wide Web[C]. West Newton, MA: User Modeling Inc, 1997. 3-12.