

Web 使用挖掘技术分析

李超锋^{1,2} 卢炎生¹

(华中科技大学计算机学院 武汉 430074)¹ (中南民族大学管理学院 武汉 430074)²

摘要 介绍了一种 Web 挖掘的分类,包括 Web 内容挖掘、Web 结构挖掘和 Web 使用挖掘。讨论了 Web 使用挖掘过程的三个步骤,即数据获取与数据预处理、模式发现和模式分析,详细分析了每一个步骤中所使用的技术。指出了目前 Web 使用挖掘研究存在的不足,给出了 Web 使用挖掘未来的研究方向。

关键词 Web 使用挖掘,数据预处理,模式发现,模式分析

Analysis of Web Usage Mining's Technique

LI Chao-Feng^{1,2} LU Yan-Sheng¹

(Computer School, Huazhong University of Science and Technology, Wuhan 430074)¹

(Management School, South Central University for Nationality, Wuhan 430074)²

Abstract This paper introduces a taxonomy of Web mining, including Web Content Mining, Web Structure Mining and Web Usage Mining; discusses three steps of Web Usage Mining, including data obtainment and data preprocess, pattern discovery and pattern analysis, analyzes the technique used in each step detailed; indicates shortage of research on Web Usage Mining at present and gives the research directions on Web Usage Mining in the future.

Keywords Web usage mining, Data preprocess, Pattern discovery, Pattern analysis

1 引言

随着每天大约发布 43 亿在线文档和 2000 万新的 Web 网页^[1],使得 Web 的信息量以惊人的速度增加。大量的信息在给人们带来方便的同时也带来了很多问题,如信息过量,难以消化;信息真假难以辨识;信息安全难以保证;信息形式不一致,难以统一处理等等。因此,人们迫切需要能自动地从 Web 上发现、抽取和过滤信息的工具,由此产生了 Web 挖掘。

Web 挖掘是从 Web 文档和 Web 活动中发现和分析潜在的、用户感兴趣的有用信息。它以从 WWW 的海量数据中挖掘有用知识为目标,将传统的数据挖掘技术与 Web 结合起来^[2-4],并综合运用了统计学、计算机网络、数据库与数据仓库、可视化等众多领域的技术。Web 挖掘分为三类:Web 内容挖掘、Web 结构挖掘和 Web 使用挖掘,如图 1 所示。

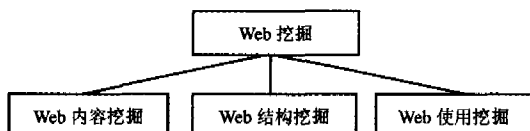


图 1 Web 挖掘的分类

Web 使用挖掘通过将数据挖掘技术应用到服务器日志及运用其它一些启发式的、Web 结构、Web 内容、user profile、注册数据等来抽取用户的浏览模式。Web 使用挖掘的结果对提供个性化服务与定制、改进 Web 系统性能和结构、改善 Web 站点结构、为商业组织提供商业智能和向用户推荐页面等方面都有重要的理论和实际意义。Web 使用挖掘也因此而成为目前国内外研究的一个热点。

2 Web 使用挖掘过程分析

Web 使用挖掘是应用数据挖掘技术从用户浏览 Web 时

交互作用的二次数据发现使用模式,从而达到更好地理解和服务基于 Web 应用需求目的的过程。Web 使用挖掘过程分三个步骤来进行^[5],数据获取与预处理、模式发现和模式分析,如图 2 所示。

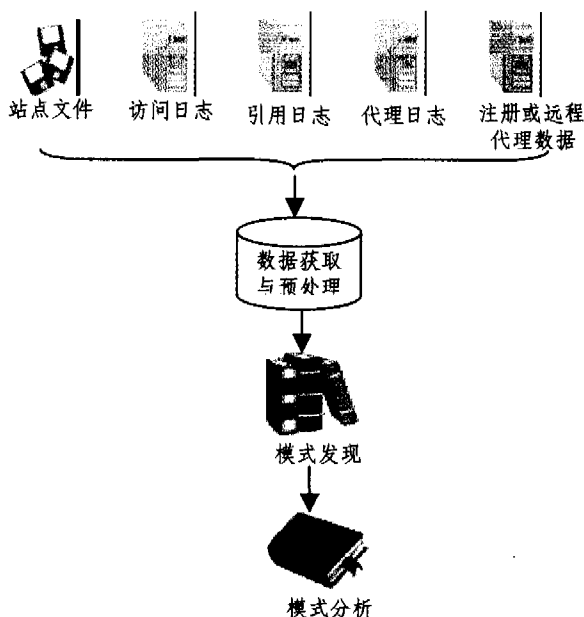


图 2 Web 使用挖掘过程

2.1 数据获取与预处理

Web 使用挖掘的数据来源于 Web 服务器与用户的交互过程中,这些必要的数据分布于服务器端、客户端和代理服务器端。数据获取就是从服务器日志、引用日志,代理服务器日志和客户端收集到这些数据。数据的收集机制与技术对于数据源的完整性来说就是至关重要的。对于这几种数据的收集来说,服务器日志数据的收集最为简单,并且其中包含了大部

分用户使用数据。但服务器日志中的数据并不全面,还有许多反映用户精确使用的数据在代理服务器和客户浏览器中,而代理服务器和客户浏览器中数据的收集却比较困难,这是因为既要考虑用户的隐私问题又要考虑 Web 服务的效率。因此数据获取过程一个主要技术关键就是如何获得全面而精确的用户使用数据,尤其是代理服务器端和客户端数据。另外,收集到的数据中可能有冗余的和遗漏的数据,因而有必要在挖掘之前对这些数据进行处理以得到适合于进行模式发现的数据格式。处理的过程一般包括数据清洗、用户识别、会话识别、事务识别和路径补充等过程。

2.1.1 数据清洗 就是把 Web 服务器日志文件中冗余的或不相关的数据去除掉,把 Web 日志转化为适合数据挖掘的可靠的精确数据。Web 日志中可能包含的冗余的或不相关的数据项有:带有“error”或“failure”状态的项;由自动搜索引擎代理生成的访问记录;带文件名后缀如 gif、jpeg、GIF、JPEG、jpg、JPG 和 map 的登录项等。

2.1.2 用户识别 用户是指使用客户交互式地检索和提供资源或显示资源的主体。用户识别是将用户和请求的页面相关联的过程。CLF 及 ECLF 仅仅记录最原始请求的主机或代理 IP,因此不同的访问人员共享同一台主机或代理服务器就不能识别出来。用户识别常用的方法有^[6~8];不同的 IP 属于不同的用户;一旦发现用户端浏览器软件或操作系统发生改变,则认为是新用户;将访问日志和引用日志与站点的拓扑结构结合构成用户的浏览路径,若发现用户正在请求页面,又不能从已访问的任何页面到达,则认为是新用户;Cookies 方法;远程代理法等。

2.1.3 会话识别 就是把每个用户在一段时间内的点击流(所有的请求页面)分解为单个的会话,最常用的方法是利用时间戳和 Cookies 方法来划分。

2.1.4 事务识别 常用于关联规则和序列模式挖掘,事务即语义相关的网页集,事务识别就是对用户会话进行语义分组,常用方法有引用长度法、最大向前引用法和时间窗法等^[6]。

2.1.5 路径补充 就是将由于本地缓存或代理服务器缓存所造成的遗漏的请求页面补充完整,目前常用的方法类似于用户标识的方法。

数据库知识发现中关键步骤之一是为数据挖掘任务建立一个合适的目标数据集。在 Web 挖掘中,数据可从服务器端、客户端、代理服务器端收集,或者从组织的数据库(包含了商业数据或经过整理的 Web 数据)获得。每一种类型的数据收集不仅在数据源位置方面不同,而且在可用数据的种类、被收集的数据属性和它的实现方面上都不相同。数据获取的任务就是全面地收集这些数据。预处理的目标是把收集到的数据源中的信息转化为适合数据挖掘和模式发现所必需的数据抽象概念,然后在事务数据库上实施挖掘算法,以期最终获得有价值的规律。预处理的结果作为数据挖掘的数据源,其质量将直接影响数据挖掘的结果,一个好的数据源,不仅可以使数据挖掘具有准确性较高的结果,还可以改进数据挖掘的算法,提高挖掘的速度。因此,Web 使用挖掘的第一步就是全面获取数据并进行预处理。

2.2 模式发现

Web 使用挖掘的目的就是要从大量的 Web 信息中挖掘出隐藏的、有价值的,对不同的用户来说有“兴趣”的信息。因此,针对不同的用户,根据预处理之后的数据特征设计出能发现有“兴趣”规则的 Web 使用挖掘算法,对 Web 使用挖掘来说就是其最基本的任务。模式发现阶段所解决的正是这一问

题。模式发现集合了数据挖掘、机器学习、统计学和模式识别等研究领域的算法和技术^[9,10]。目前常用的模式发现规则包括统计分析、关联规则、聚类、分类、序列模式和依赖建模等。

2.2.1 统计分析 是抽取有关网站访问者信息的最常用的方法。通过分析会话文件或事务数据库,可对诸如网页视图、浏览时间、导航路径长度等做出不同种类的描述性统计分析。尽管这种分析缺乏深度,但这类知识有助于改进系统性能、提高系统的安全性、便于站点修改,并能提供营销决策支持^[5]。

2.2.2 关联规则 关联规则挖掘就是挖掘出用户在一个会话期间从服务器上访问的页面/文件之间的关系,找出在某次服务器会话中最经常一起出现的相关页面。除了商业和市场营销方面的应用之外,这类规则还有助于网站设计者重新组织和设计 Web 站点的结构。

2.2.3 序列模式 序列模式挖掘就是挖掘出交易集之间有时间序列关系的模式,在 Web 日志中发现所有满足用户规定的最小支持度的大序列模式。其结果有助于商家印证其产品所处的生命周期阶段,增加广告的针对性;有助于网管组织网站结构等。

2.2.4 分类 分类技术主要是根据用户群的特征挖掘用户群的访问特征(某些共同的特性),其结果可用于分析访问某一服务器文件的用户特征,从而可用于页面推荐等。

2.2.5 聚类 聚类技术是对符合某一访问规律特征的用户进行用户特征挖掘,其结果对电子商务和为用户提供个性化的服务特别有用。

2.2.6 依赖性建模 其目标是开发出一种能表达 Web 领域中各种变量之间的显著依赖性的模型,其结果对设计增加网上产品销售策略以及改进用户导航的便利性都很有用。

Web 使用挖掘的算法虽然是利用传统的数据挖掘算法,但数据源和具体的应用环境发生了根本的变化。从这个意义上来说,Web 使用挖掘无疑是一个全新的研究领域。因此,这一阶段的技术关键是设计出高效的挖掘算法。其解决方案可以在借鉴传统的数据挖掘算法的基础上,结合 Web 使用挖掘数据源中数据分布性的特征及具体的挖掘应用环境来设计算法,同时要通过试验来验证所设计算法的质量与效率。

2.3 模式分析

模式分析是 Web 使用挖掘的最后阶段。这个阶段的目标是从模式发现过程的输出中去除不相关的规则或模式及抽取有兴趣的规则或模式。目前研究的常用的模式分析方法和工具包括可视化、OLAP 和类 SQL 查询机制等。

2.3.1 可视化 已经非常成功地用于帮助人们理解各种现象,无论是客观存在的还是抽象的。因此用它来理解 Web 用户的行为就是自然的选择。Pitkow 等人已经开发出了 WebViz 系统来可视化 WWW 访问模式^[11]。该系统提出了在一个 Web 路径聚合体中服务器登录项被用于抽取称为 Web 路径的 Web 访问模式结果。WebViz 允许分析人员有选择性地分析感兴趣的 Web 部分而过滤掉不相关的部分。Web 被可视化为一个带环有向图,其中节点是网页而边是(网页之间的)超链接。

2.3.2 在线分析处理(OLAP) 是作为商务处理中的数据库策略分析的一个非常强大的范例而出现的。Web 使用数据的分析需要同数据仓库非常相似,因此 OLAP 技术也非常适用。目前在这方面的研究是把挖掘结果放入数据立方体中来实现。

2.3.3 查询机制 其实现主要是结合结构化查询语言 SQL 来实现。许多 Web 使用分析工具,比如 WUM^[12]、Web-

Miner^[5]、Midas^[13],首先提供一些对象规则,比如支持度和可信度,这些规则有助于手工过滤一些被认为不重要的知识,然后利用类 SQL 语句得到分析结果。

Web 使用挖掘算法挖掘出的模式中可能包括对用户来说不感兴趣、冗余的模式。此外,挖掘出的结果可能是抽象的。模式分析要解决的技术关键就是过滤掉用户不感兴趣的模式,同时以一种易于用户理解的形式把挖掘的结果展示给用户。这就要求有集统计分析、可视化分析技术、过滤、解释功能等为一体的工具。不仅能过滤已经发现的访问模式,还要能解释挖掘出来的知识,以辅助人们理解用户的访问模式,而且还要具有有关特定问题的知识,能指出问题的所在。解决方案是把类 SQL 查询机制、信息过滤机制、OLAP 分析技术和可视化技术集成到一起,共同向用户提供一个友好的人机界面。

3 Web 使用挖掘研究存在的不足及未来的研究方向

Web 使用挖掘的研究目前还是一个全新的研究领域,国内外研究人员虽然进行了一些研究,但总的说来不存在以下一些不足或尚须改进的地方。

3.1 在数据获取与预处理方面

Web 使用挖掘结果的质量首先取决于数据获取的完整性与预处理结果的质量,但由于 Web 数据的分布性和异质性使得这一问题变得相当困难,表现在以下几个方面:

1. 数据获取的矛盾

Web 领域中的一个问题是分析人员的分析需要(要求收集的使用数据更详细)和用户的保密性需要(要求个人的数据尽可能少地被收集)之间存在固有的冲突。目前的研究中尚没有这一问题的完美解决方案,有的研究系统忽略了这一问题。

2. 数据集成

Web 使用挖掘所需要的数据分布于 Web 服务器日志、引用日志、代理服务日志、注册文件和索引服务器日志中。来自于这些不同形式的数据源中的数据的智能集成性和信息相关性能够揭示使用信息,但这些信息不可能显式地来源于以上数据源中任何一个。目前对于如何较好地集成这些数据还有待于进一步研究。

3. 事务识别

从各种日志中收集的 Web 使用数据具有非常好的粒度。但是,这些数据在具有极度的通用和相当详细的优点的同时,还具有不能直接进行分析的相应缺点,因为用户的分析可能一开始侧重于微观趋势而不是宏观趋势上。另一方面,究竟侧重点在微观趋势还是宏观趋势上依赖于特定的分析目的。因此,有必要在把它们送进挖掘系统之前把独立的数据集事件组合或分解为 Web 事务。而目前所开发的系统在这一方面所做的工作还远远不够。

3.2 Web 使用挖掘过程

Web 挖掘采用了来自多个领域的技术和先验知识,又有自身的一些技术,但仍有许多新技术需要研究,特别是以下问题:

1. 新型的知识

Web 使用挖掘的结果通常依赖于关联规则、序列模式等常用的挖掘方法,随着人们使用 Web 方式的增加,必定会出现新的使用模式,因此,需要研究有关于用户使用模式的新知识。

2. 挖掘算法性能的提升

目前的挖掘工具不多,现有的挖掘工具功能都有较大的局限性,所采用的算法质量也不高。算法的质量直接影响挖掘出来的知识的有效性,也影响系统的挖掘效率。因此,算法效率的提高也是一个长期的研究课题。

3. 分布式、多媒体 Web 数据的挖掘技术研究

如何尽可能全面(不可能全部)地集成 Web 上动态、分布的数据并实施挖掘,以挖掘出更多有用信息;又如何将不同数据源挖掘出来的信息进行集成,即多种数据(原始或挖掘出来)的智能集成系统的研究与开发,以期能提供完善的查询、优化和维护机制是 Web 使用挖掘的又一重要研究领域和方向。

3.3 模式分析工具的研究

Web 使用挖掘算法的输出经常不以直接适合于人们接受的方式出现,因此就有必要开发帮助分析人员更好地消化知识的技术和工具。下面是一些在这个领域中有必要提及的一些问题:

1. 使用分析工具

有必要开发把统计方法、可视化和人员因素结合起来以帮助理解已挖掘知识的工具。

2. 已挖掘知识的解释

开发出来的使用分析工具需要与相关特殊问题领域的特定知识相结合,而不仅仅只考虑已发现规则和模式的统计属性。

结束语 随着 Web 上的信息量的快速增长,隐藏在 Web 中的用户浏览模式将对 Web 浏览用户获取有兴趣的信息、Web 站点设计人员更合理地设计 Web 站点和商业组织获得高质量的市场决策信息产生重大的影响。因而,对 Web 使用挖掘的研究也必将成为一个重要的研究方向。本文讨论了一种 Web 挖掘的分类,详细分析了 Web 使用挖掘的过程,给出了 Web 使用挖掘目前研究的不足并提出了未来的一些研究方向。

参考文献

- 1 Tanasa D, Trousse B. Advanced data preprocessing for intersites Web usage mining [J]. IEEE, 2004, 19(2): 59~65
- 2 Daganl F R. Knowledge discovery in textual databases (KDT) [J]. In: Proc. of 1st Int'l Conf on Knowledge Discovery and Data Mining, Montreal, Canada, 1995, 112~117
- 3 Chakrabarti S. Data mining for hypertext: a tutorial survey [J]. SIGKDD Exploration, 2000, 1(2): 1~11
- 4 Cooley R, Mobasher B, Srivastava J. Web mining: Information and pattern discovery on the World Wide Web [C]. In: Proc. of the 9th Int'l Conf. on Tools with Artificial Intelligence (ICTAI'97), CA, 1997, 558~567
- 5 Srivastava J, Cooley R, Deshpande M, Tan Pang-Ning. Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data [J]. SIGKDD Explorations, 2000, 1(2): 1~12
- 6 Cooley R, Mobasher B, Srivastava J. Data Preparation for Mining World Wide Web Browsing Patterns [J]. Journal of Knowledge and Information System, 1999, 1(1): 5~32
- 7 Pitkow J. Insearchof reliable usage data on the WWW [C]. In: Sixth International World Wide Web Conference, Santa Clara, CA, 1997, 451~463
- 8 Shahabi C, Zarkesh A, Adibi J, Shah V. Knowledge discovery from users Web-page navigation [C]. In: Workshop on Research Issues in Data Engineering, Birmingham, England, 1997
- 9 Chen M S, Han J, Yu P S. Data mining: An overview from a database perspective [J]. IEEE Transactions on Knowledge and Data Engineering, 1996, 8(6): 866~883
- 10 Fayyad U, Piatetsky-Shapiro G, Smyth P. From data mining to knowledge discovery: An overview [J]. In: Proc. ACM KDD, 1994
- 11 Pitkow J, Bharat K K. WebViz: A tool for world-wide web Access log analyses [J]. In: First Intl. WWW Conf. 1994
- 12 Spiliopoulou M, Faulstich L C. WUM: A Tool for Web Utilization Analysis [C]. In: EDBT Workshop WebDB'98, Valencia, Spain, Mar. 1998
- 13 Spiliopoulou M, Faulstich L C, Winkler K A. Data Miner analyzing the Navigational Behaviour of Web Users [C]. In: Proc. of the Workshop on Machine learning in User Modeling of the ACAI'99 int. Conf., Creta, Greece, July 1999