WEB 使用挖掘研究

Review of the Research for Web Usage Mining

黄 浩 （北京航空航天大学 北京 100083、 北京经济管理职业学院 北京 100102）
王建军 （大连理工大学 大连 116023）

摘 要： 本文介绍了 Web 使用挖掘的理论背景、基本原理及其主要的应用领域。以 Web 使用挖掘的过程为框架，综述了 Web 使用挖掘在挖掘方法和应用方面的研究成果。分析了 Web 使用挖掘的研究难点和未来的主要研究领域。

关键词： Web 使用挖掘 Web 挖掘 模式识别 Web 应用

随着 Internet 的快速发展，电子商务已经成为企业从事商业活动的一种重要方式。相对于传统商务模式，电子商务领域的竞争更加激烈，客户只需点击鼠标就可以在不同的商家之间转换、比较。因此，网站的布局、用词、服务等任何一个细节都可能成为吸引或失去用户的关键因素。

如何充分了解网络客户的个性化需求以及网络客户的浏览行为，成为电子商务网站结构设计、内容设计和服务设计的前提条件。网络交易可以生成大量的交易记录，同时，客户的浏览行为也有意或无意地在网上留下了“痕迹”。通过数据挖掘的技术，对所获得的各种 WEB 信息（包括交易记录、Web Log、注册信息、cookies 信息、所访问站点的结构与页面信息等）进行 WEB 使用挖掘，发现客户的访问模式，从而可以对客户进行分类、聚类、发现潜在的客户、改进站点的设计，方便客户的浏览和交易，并为客户提供个性化的服务。可以看出，WEB 使用挖掘研究的意义在于它是推动电子商务向智能化、个性化发展的重要动力。

1 Web 挖掘

Web 挖掘就是从与 WWW 相关的资料和行为中抽取感兴趣的、有用的模式和隐含信息[1]。Web 挖掘将传统的数据挖掘技术与 Web 结合起来，它可以在搜索引擎的开发和改进、确定权威页面、Web 文档分类、Web Log 挖掘、智能查询等方面发挥积极地作用。

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 使用挖掘（web usage mining）。

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。这些数据既有来自数据库的结构化数据，也有用 HTML 标记的非结构化或半结构化数据。根据其使用的方法，Web 内容挖掘可分为信息查询和数据库两种方法[2]。根据其挖掘策略的不同又可分为 Web 概要（直接挖掘 Web 文档的内容）和搜索引擎结果概要（对搜索引擎的查询结果做进一步处理，得到更精确和有用的信息）两种方法。

1.2 Web 结构挖掘

Web 结构挖掘是从 WWW 的组织结构、Web 文档结构及其链接关系中发现知识。由于 Web 文档之间的是以超链接形式组织的，因此，WWW 不仅能够提供 Web 文档的内容信息，同时也揭示了文档之间的关联关系。通过对 Web 结构的分析，可以发现页面结构和链接关系中所蕴含的有用模式；也可以对页面及其链接进行分类和聚类，发现权威页面。有关这方面的算法研究成果有：Page – rank、HITS（Hyperlink – Induced Topic Search）及改进的 HITS（将内容信息加入到链接结构中去）、Hub/authority（Kleinberg，1998）[3]。

1.3 Web 使用挖掘

Web 使用挖掘是应用数据挖掘的技术从 Web 数据中发现用户访问模式的过程[4]。它可以帮助我们提高 Internet 信息服务的质量，改进 Web 服务器的系统性能和结构。

©《计算机系统应用》编辑部 http://www.c-s-a.org.cn

2　Web 使用挖掘的流程

Web 使用挖掘总体上可以分为三个阶段：数据预处理、模式发现和模式分析。

2.1　数据预处理

由于网络服务器日志并非专门用于数据挖掘，因此在进行 Web 使用挖掘之前，必须对"杂质"数据进行过滤，比如消除数据中的不一致性、将多个数据源中的数据统一为一个数据存储等。预处理的结果会直接影响到挖掘算法产生的规则和模式，因此数据预处理的效果是 Web 使用挖掘质量的保证。数据预处理主要包括站点识别、数据选择、数据净化、用户识别和会话识别等五个关键问题。

站点识别将产生网站结构图，通过网站结构图可以抽取和过滤浏览页面，有助于最终识别会话。模式分析中也需要参考网站结构图来分析产生的模式。

数据选择是从网络日志文件中选择所需的字段。

数据净化是删除 Web 服务器日志中与挖掘算法无关的数据，这类数据通常有三种：图片、框架等非用户请求逻辑单位；Web Robot 的浏览日志记录；噪音和错误信息。

由于代理服务器、防火墙、Internet 服务提供商采用动态分配 IP 地址等问题，使得用户识别变得复杂。目前常用的一些识别用户的方法包括：IP 地址和代理（agent）；嵌入 sessionID；cookie 和注册等方法。

会话识别就是将用户的访问记录分为单个会话。有四种识别会话的模型：页面类型模型（Page Type Model）、参引长度模型（Reference Length Model）、最大前向参引模型（Maximal Forward Reference Model）和时间窗口模型（Time Window Model）。常采用的是时间窗口模型，即假定用户一次访问的时间有最大限制，以用户访问历时作为划分会话的分界。根据 Perkowitz 统计的结果[5]，一般间隔时间为 25.5 分钟。最大前向参引法是从用户访问的首页开始，到第一个回退动作作为一次会话，接下来的第一个向前动作引发下一个会话，直到下一个回退动作产生。这样，可将用户访问页面序列划分为不同的会话。

2.2　模式发现

模式发现就是利用挖掘算法挖掘出有效的、新颖的、潜在的、有用的及最终可以理解的信息和知识。模式的发现可以应用多个领域的方法，如统计学、机器学习、模式识别等。常用的模式发现技术包括：路径分析（Path Analysis）、关联规则挖掘、时序模式发现、聚类和分类等技术。

路径分析可以用来发现 Web 站点中访问最频繁的路径，从而可以调整站点结构。该技术常用的方法是图，通过网站的结构图，把网站页面定义为节点，页面之间的链接关系定义为图中的边，利用图论的理论和实践进行分析。文献[6]设计了网站的访问矩阵，通过对该矩阵进行计算得到网站的偏爱子路径。

关联规则挖掘可用于发现用户之间、页面之间或用户浏览页面和网上行为之间的潜在关系。例如通过关联挖掘可能得出"浏览页面 A 的用户 70% 都将浏览页面 B"或"浏览页面 C 的用户，60% 都会下订单"等规则。目前，在关联规则挖掘算法方面的研究已经很深入，大量算法被提出，如 Apriori 算法、FP 树算法、A－Close 算法等。

时序模式发现的目的是从用户访问序列中挖掘出相关规则。Web 服务器事务日志中记录的是一段时间内用户的访问行为，在数据预处理阶段，每个事务都会附带时间戳。通过对 Web 使用数据的时间序列分析，可以预测用户的访问行为。

分类分析是为具有某些公共属性的特定群体建立概要特征。例如：在页面 D 进行过在线定购的客户中 80% 是 15－35 岁的城市人。得到这种分类后就可以针对这类客户的特点开展网上商务活动。目前，最为典型的分类方法是基于决策树的分类方法，如 ID3，它采用自顶向下不回溯策略，确保找到一个简单的树。算法 C4.5 和 C5.0 都是 ID3 的扩展。

聚类是将数据对象按特征相近的原则划分为多个类或簇。例如：一些用户最近经常访问页面 E，经过分析，把这些用户聚为一类，他们的特征是即将结婚。聚类和分类是一个互逆的过程。Web 使用挖掘中有两种有特点的聚类：使用聚类和页面聚类。使用聚类就是将经常访问相同页面的用户群区分出来，对他们开展特定的广告策略或个性化服务。页面聚类是发现内容相关的页面组，为搜索引擎和 Web 服务商提供相关信息。Web 使用挖掘中，聚类算法将用户浏览页面看作一个数据空间，构造稀疏图。根据每个页面内容的相似性和路径的关联性，将对象分割为若干 k－最近邻

©《计算机系统应用》编辑部 http://www.c-s-a.org.cn

居子图,图中每一个点表示一个页面,子图的密度作为边的权重被记录下来。当发现两个子图的关联性或相似性与子图内部页面相关性较强的话,则合并两个子图,直到聚类最终结束。文献[7]提出了基于竞争的自激励神经网络学习算法 SIN,并将其运用于 Web 日志挖掘,实现用户聚类和页面聚类。文献[8]提出了在多维空间中基于概念层次模型的 Web 使用数据聚类的方法。

2.3 模式分析

使用各种技术挖掘出来的模式,需要合适的工具和技术对其进行分析、解释、可视化,只有这样,挖掘出来的各种模式才能被有效地利用。因此,模式分析的目的是通过用户的选择和观察,把发现的规则模式转换为知识。模式分析技术和工具也是近年来 Web 使用挖掘研究的热点之一。常用的模式分析方式有:类似于 SQL 的查询机制、联机分析处理技术 OLAP;可视化技术。

类似于 SQL 的查询机制能够自动搜索相关的规则、模式,帮助分析用户的目的,以类似于 SQL 的智能的方式回答用户的查询。它不仅可以直接给出用户所指定的属性,还可以提供辅助决策的附加信息。目前的研究已经在 SQL 语言的基础上提出了几种适合在数据挖掘过程中使用的查询语言,如 DMQL,也有为 Web 挖掘而专门定义的查询语言,如 WebSSQL、WebLQM 和 Squeal 等。

联机分析处理技术 OLAP(On – Line Analytical Processing)是在基于多维数据模型的数据仓库上使用的分析技术,它通过上卷、下钻、切片和切块、旋转等操作,实现多维数据环境下的知识发现。Web 服务器上的数据呈海量增长趋势,且具有很强的时间特性,因此,Web 使用数据的分析需要数据仓库的支持。OLAP 允许基于主题对数据进行查询和分析,使分析人员能够对信息进行快速、一致、交互地存取,这些与可视化工具的结合将增强 Web 使用挖掘的能力和灵活性。这一领域的研究已得到广泛关注,有待于进一步的发展。

可视化技术用图形和图像表示抽象复杂的关系,用文字描述模式的内涵,可以帮助人们更好地理解 Web 中大量复杂数据之间的联系。Pitkow 等人已经开发了 WebViz 系统将 Web 的访问模式可视化。IDL(Interactive Data Language)交互数据语言是第四代可视化语言,它支持 OpenGL 图形加速、量化可视化表现、集成数学与统计学算法、连接 ODBC 兼容数据库和多种程序连接工具等,是目前科学数据可视化方面较好的工具。

3 Web 使用挖掘的应用

通过 Web 使用挖掘分析,可以:(1)发现用户的访问模式信息,理解用户的行为和意图,从而可以为用户提供个性化的服务;(2)了解 Web 结构设计的实际效果,检测系统的性能,改进 Web 站点的结构和服务水平,使得 Web 站点能随着时间、需求的变化而不断调整。

3.1 个性化推荐

Web 个性化推荐系统通过收集和分析用户信息来学习用户的兴趣和行为,对用户可能访问的网页进行预测,从而实现主动推荐的目的,也称为个性化服务(Personalization)[9]。在日趋激烈的电子商务环境中,个性化推荐系统能发现潜在的客户,保留现有客户,提高销售能力。许多大型电子商务系统都不同程度的使用了各种形式的推荐系统,如 CDNOW、Ebay、dangdang 等。

个性化应用的基本原理是过滤掉与用户行为无关的信息,使用户与可能感兴趣的资源匹配。信息过滤技术分为基于内容的过滤(Content – based Filtering)和协同过滤(Collaborative Filtering)。基于内容过滤的技术是通过比较用户档案(User Profile)和资源的属性来推荐资源。它通过相似度计算(就矢量空间而言,通常采用余弦度量的方法)实现二者的比较,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。协同过滤是根据用户之间的相似性来推荐资源。它比较的是不同的用户档案,而不是资源属性和用户档案。它的主要原理是聚类,即通过为当前用户寻找 k 个最相似的邻居来预测用户的个性化需求。它的主要问题是稀疏性问题,即系统使用初期由于不能获得足够多的用户,而很难利用已有的基准对新的用户进行聚类分析。

文献[10]提出了推荐系统在电子商务领域内的机遇和挑战。文献[11]是对个性化推荐算法的探索性研究,文献[12]提出了频繁路径集的挖掘浏览模式在个性

©《计算机系统应用》编辑部 http://www.c-s-a.org.cn

化服务中的应用。

3.2 Web 系统优化

对于网络用户而言，Web 站点的性能和服务质量是至关重要的。Web 使用挖掘能够发现用户浏览网站的行为习惯，因此，Web 系统的管理者可以在网站结构的安排、Web 缓存等方面对 Web 系统进行优化。根据发现的用户实际浏览情况，调整网页的链接和内容。文献[13][14]中，Perkowitz 提出了自适应网站的概念，并对其进行了研究。自适应网站是能根据用户访问模式自动或半自动地学习和调整自身组织架构的一种网站，它是以 Web 使用挖掘为基础的。文献[15]利用不断追踪和挖掘网络用户行为的技术，提出了网站设计的 GIST 模型，它能够指导商业网站以客户为核心，不断自我调整，是 Web 使用挖掘在设计科学领域的研究。

另外，Web 使用挖掘还可以被广泛应用于客户关系管理、客户信用分析和欺诈甄别等方面。

4 Web 使用挖掘研究难点及主要研究领域

4.1 Web 使用挖掘研究的难点

Web 使用挖掘的核心技术基础是数据挖掘的技术，因此，数据挖掘技术的发展直接决定着 Web 使用挖掘的发展。另外，根据 Web 使用挖掘的自身特点，其研究难点主要包括：(1) Sever log 提供的信息太少，由于无法获得更多的数据，给 Web 使用挖掘带来了难度；(2) 动态网页的大量使用使得分析 Log 变得更为困难；(3) Session 的分析一直是难点，文献[16]在这方面进行了研究；(4) Crawlers 的过滤；(5) 巨大的数据量及自动转换；(6) 市场层面的洞察力。

4.2 Web 使用挖掘研究的主要领域

Web 使用挖掘研究领域主要包括：(1) 个性化推荐系统，这其中还可以分为几个重要的研究点，包括如何结合人口统计信息、购买信息和用户反馈信息建立推荐系统；如何把推荐系统和市场决策分析结合起来；用户数据共享的研究；Web 使用挖掘中道德问题的研究；(2) Web 使用挖掘技术与现有的电子商务系统的集成研究；(3) 改造和构造新的 Web 使用挖掘算法的研究，文献[17][18]中把用户的浏览模式和购买模式结合起来考虑，构建了一个算法用以挖掘用户的交易模式（使用规则来表示）；(4) 利用 Web 使用挖掘技术，实现 Web 站点优化的研究。

5 结束语

Web 挖掘是利用数据挖掘技术对 Web 上的信息、结构进行的挖掘，Web 使用挖掘是 Web 挖掘的一个分支，它主要是通过分析用户浏览行为所留下的信息，发现客户的特征、偏好。通过 Web 使用挖掘，可以实现个性化推荐和网站的调整、优化。Web 使用挖掘是一个较新的研究领域，有许多问题有待于进一步研究，它的研究成果也将推动电子商务应用的发展。

参考文献

1 韩家炜、孟小峰、王静，Web 挖掘研究[J]，计算机研究与发展，2001，38(4)：405－413.

2 Zainane O R et al. Multimediaminer：a system prototype for multimedia data mining[C]. In：proc ACM SIGMOD Int conf On management of Data，1998：581－583.

3 Wang K，Zhou S，Liew S C. Building hierarchical classifiers using class proximity[C]. In：proc of VLDB'97，Edinburgh，UK，1999：363－374.

4 Jaideep Srivastava，Robert Cooley，Mukund Deshpande，Pang－Ning Tan，Web Usage Mining：Discovery and Applications of Usage Patterns from Web Data[J]. SIGKDD Explorations，Vol. 1，Issue 2，2000.

5 J. Pitkow. Summary of www Characterization －s[C]. In：7th International World Wide Web Conference，1998.

6 邢东山、沈钧毅、宋擒豹，从 Web 日志中挖掘用户浏览偏爱路径[J]，计算机学报，2003，26(11)：1518－1523.

7 董一鸿、庄越挺，基于新型的竞争型神经网络的 Web 日志挖掘[J]，计算机研究与发展，2003，40(5)：661－667.

8 Supriya Kumar D E，Radha Krishna P. Mining Web data using clustering technique for Web personalization[J]. International Journal of Computational Intelligence and Applications，2002，2(3)：255－265.

9 Mobasher B，Cooley R，Srivastava J. Automatic Personalization Based on Web Usage Mining[J]. Communications of the ACM，2000，43(8)：142－151.

（下转第 124 页）

©《计算机系统应用》编辑部 http://www.c-s-a.org.cn

(上接第128页)

10 J. Ben Schafer, Joseph A. Konstan, John Riedi, Recommender systems in e-commerce [C]. Proceedings of the First ACM Conference on Electronic Commerce (EC-99), November 3-5, 1999, Denver, CO, USA. ACM, 1999 158-166.

11 曾志聪、姚国祥，基于Web挖掘的个性化系统[J]，计算机工程与设计，2006，27(7)：1155-1157.

12 丁一、卢正鼎，基于Web挖掘的用户服务研究[J]，计算机仿真，2004，21(6)：83-84.

13 Mike Perkowitz, Oren Etzioni: Towards adaptive Web sites: Conceptual framework and case study[J]. Artificial Intelligence 118(1-2): 245-275 (2000).

14 Mike Perkowitz, Oren Etzioni: Adaptive Web Sites: an AI Challenge[J]. IJCAI (1) 1997: 16-23.

15 Terri C. Albert, Paulo B. goes, Alok Gupta: GIST: A Model for Desigan and Management of Content and Interactivity of Customer-Centric Web Sites[J]. MIS Quarterly vol. 28 No2, 161-182/June 2004.

16 B. Berent, M. Spiliopoulou, J. Wiltshire. Measuring the Accuracy of Sessionizers for Web Usage Analysis [C]. Proceedings of the Web Mining Workshop at the First SIAM International Conference on Data Mining, April 2001, Chicago.

17 C.-H. Yun and M.-S. Chen. Mining Web Transaction Patterns in an Electronic Commerce Environment[C]. Proc. of the 4th Pacific-Asia Conf. on Knowledge Discovery and Data Mining, pp. 216-219, April 18-20, 2000.

18 C.-H. Yun and M.-S. Chen, Using Pattern-Join and Purchase-Combination for Mining Web Transaction Patterns in an Electronic Commerce Environment [C]. Proc. of the 24th annual Intern1 Computer Software and Application Conference (COMPSAC-2000), pp. 99-104, October 25-27, 2000.

©《计算机系统应用》编辑部 http://www.c-s-a.org.cn