网站结构和内容对 Web 使用挖掘的影响*)

刘丽珍[1] 宋瀚涛[1] 陆玉昌[2]

(北京理工大学 北京100081)[1] (清华大学 北京100084)[2]

Web Usage Mining Process Influenced by Web Site Structure and Content

LIU Li-Zhen[1] SONG Han-Tao[1] LU Yu-Chang[2]

(Beijing University of Science and Technology, Beijing 100081)[1] (Tsinghua Univesity, Beijing 100084)[2]

Abstract The Paper emphasizes relativity between Web usage mining and the application of Web site structure and content. It has shown that the amount of effort involved in processing and quantifying the structure and content of a Web site is well worth in performing Web usage mining. The necessity of combining Web site structure and content with Web usage mining process is further proved.

Keywords Web usage mining, Page view, Page file

Web 使用挖掘的应用日益广泛，尤其是在电子商务的大力支持下，越发显示出蓬勃的生命力。它通过数据挖掘技术对 Web 上的数据进行挖掘，从而发现 Web 上的用户使用模式。但 Web 使用挖掘的成功与否和网站的结构设计和内容安排有着密不可分的联系，反过来，Web 使用挖掘的结果又能服务于网站结构和内容的设计与安排，二者相辅相成。

1 网站内容的处理

虽然网站的超链结构已经自然形成了一个很直接的拓扑图，但是页面文件中内容的量化依然是不直观的，甚至连页面访问的静态集成、URI 的转换处理和结构对应关系也是不明朗的。

URI 转换是一个参照页面浏览的重要的处理过程。对于静态的网站，URI 转换相对来说比较容易，因为在 URI 和内容之间有个一一对应的关系。但是对于动态网站，几个 URI 组成一个内容单位，或几个内容单位对应同一个 URI，处理起来就不那么容易了。

对于多个 URI 对应一个内容单位是一个隐藏会话的标识，为了便于跟踪会话，一个会话 ID 常常隐藏在动态 URIs 中。例如：当引用信息隐藏在一个 URI 中，同一个内容单位将有多个 URI 对应每一个到该内容的链接。单一内容多 URI 问题通过一个有序的一系列有规律的表达的应用来解决。这个一系列的规律的表达指明一个 URI 的哪些部分负责识别内容，哪些部分是用来进行会话跟踪或用户识别等用途的。但一个网站往往需要许多这样的规律表达才能将所有可能的 URI 正确转换。

一个 URI 对应几个内容的问题是无法用 Web 服务器日志来解决的，动态网页产生于隐藏的 Post 请求或每个用户会话的静态内容服务器中，所以访问数据要从多个 Web 服务器上收集。

除了 URI 转换，内容分级组织、页面分类和页面聚类对于模式发现和模式分析的过滤输入都非常有用。一旦 URI 转换完成，我们常常可以利用内容的分级组织来了解产品的分级组织。因为网站没有分级组织，这些技术对于 Web 使用挖掘中的基于页面文件的分类和聚类是非常有用的。而且对于图形和其它多媒体文件的分类聚类也有相当的作用。为了充分利用数据挖掘算法，页面文件的文本必须要预处理。经过预处理的文本就可以利用分类和聚类进行挖掘了。

页面文件的量化问题等同于非结构文本文件的量化，一些实验试图利用 HTML 的显示标记来推理其语义含义，这在通常意义上是行不通的。

2 网站结构的处理

网站结构是由浏览页面之间的超链、框架和构成页间浏览的图像标志构成的。如果没有网站结构，那么 Web 使用挖掘预处理就无法完成。另外，网站结构对于识别潜在的有趣的规则也非常有用，而且还用在页面浏览识别和用户识别中。

由于框架的存在，网站的潜在页面浏览数量是巨大的。通常网站的每个页面浏览由两到三个框架构成，框架类型有：顶层框架、导航框架和主框架。如果一个网站有 M 个顶层框架，N 个导航框架，P 个主框架，则页面浏览的数目为 M×N×P。随着框架数的增加，页面浏览的数目将是惊人的。因此，网站结构需要被存储成框架集 F、链接和目标列表。一个目标就是一个区域，该链接应该装载到浏览显示中。对于进一步复杂的情形，一个单一的链接会导致在一个页面浏览中一个或所有框架的替换。在 Web 使用挖掘中一个网站的正式定义如下：G 是网站的有向图，h_i 是 Html 文件，l 是链接类型，o_i 是目标区。

$$G=[\langle F_1,\cdots,F_n\rangle] \quad (1)$$

$$F=\{h_f,k_1,\cdots,k_m\} \quad (2)$$

$$K=\langle l,(h_1,o_1)|\cdots|(h_p,o_p)\rangle \quad (3)$$

链接类型是指从 Web 服务器中如何请求页面文件，最普通的就是 Get，其它的方法还有 Post，Hidden Post，Ftp 和 Mail。Post 和 Hidden Post 这两种类型是用 Post Http 方法将

*)基金项目：973国家重点基础研究项目(G1998030414)。刘丽珍 副教授，博士研究生，主要研究方向为网络挖掘。宋瀚涛 教授，博士生导师，主要从事多媒体与信息管理技术、网络通信技术等的研究。陆玉昌 教授，从事 WWW 上的数据集成、数据仓储及知识发现的有效算法与软件系统等。

数据送回 Web 服务器。尽管 Post 方法理论上是一种将数据从客户端传送到服务器的方法，事实上，内容服务器的应用是将页面文件送到客户端作为在 Post 中传送数据的响应。在一个常规 Post 和一个 Hidden Post 之间的区别是客户端数据如何被送回 Web 服务器。一个常规 Post 以 CGI 格式追加数据到 URI 中，而一个 Hidden Post 是用 Http 标题传递数据，这是它们的主要区别。因为 URI 是作为 CLF 或 ECLF 的格式的一部分记录在日志中的，但是 Hidden Post 数据则不同。假如 Web 使用挖掘数据源包含了 Post 参数，那么在 Post 和 Hidden Post 之间就没有明显的区分了。

框架类型参照了 Html 框架标志的应用，在这种情况下，排列成框架的页面浏览能自动地从 Web 中请求。图1是一个简单的网站结构例子。该网站是一个树型结构，每个页面浏览是由一或两个框架组成，框架区域又分成顶层框架区、导航框架区和主框架区。

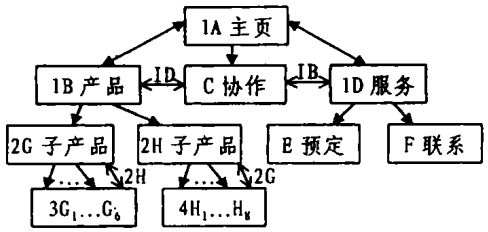

图1　网站结构示例

这个网站的结构图自然形成了一个面向对象的图，网站对象包含了一系列的框架对象，每个框架对象组成一个文件，再加上一系列的连接对象。为了从网站的图形中形成页面浏览，一个初始的页面浏览需要被指定。

G=[{index,(frame,1,left|frame,A,main)};
{1,(get,A,main),(get,B,main),(get,D,main)};
{2,(get,G,main),(get,H,main)};
{3,(get,G_1,main),…,(get,G_6,main)};
{4,(get,H_1,main),…,(get,H_8,main)};
{A,(get,C,top)};
{B,(get,2,left\G,main),(get,2,left\H,main)};
{D,(get,E,top),(get,F,top)};
{C};
{G,(get,3,left\$G_1$,main),…(get,3,left\$G_6$,main)};
{H,(get,4,left\$H_1$,main),…,(get,4,left\$H_8$,main)};
{G_1};…;{G_6};{H_1};…;{H_8};{E};{F};]

在以上的例子中，初始浏览是主页1A，网站的所有其它页面浏览的形成是被一个或多个框架区域形成，并且还有许多的候选链接。例如：页面浏览1B 是由从 Frame 1中到 B 的链接形成的；页面浏览2H 是由 B 中的单链接形成的。该图中有一个顶层框架、4个导航框架和19个主框架，所以从理论上讲它的页面浏览数最大是1×4×19＝76。但是在这个简单的例子中，每一个主框架都有一个导航框架可以达到，有19个页面浏览是由主框架和导航框架联合相成，而只有3个单独的框架页面浏览，因而该网站实际有22个可能的页面浏览。

3　网站的结构和内容对使用挖掘的影响

目前我们有许多工具可以进行数据的清理、Web 服务器日志中的会话识别，还有大量的数据挖掘算法从预处理后的数据集中发现用户使用模式和预测趋势，但最终 Web 使用挖掘的效果依然不能令人非常满意。其中一个重要的原因就是人们忽视了对使用挖掘效果起着重要影响作用的网站结构和内容。如图2所示的是 Web 使用挖掘的过程，从中不难看出，网站的结构和内容对整个 Web 使用挖掘过程的每个重要阶段都是关键性的数据源。

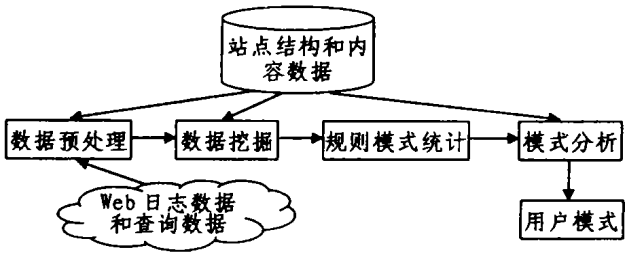

图2　Web 使用挖掘

在 Web 上有三种数据：内容数据、结构数据和用户使用数据。内容数据是指网页上实际存在的数据，是供网上用户使用的，通常是由文本和图像组成的；结构数据是用来组织内容的一种描述性的数据，主要是指页与页之间的超链接，包括页面内的 HTML 或 XML 标记的安排；而用户使用数据是指 Web 页面的使用模式，比如：IP 地址、页面引用和访问时间等数据。使用数据通常源于普通和扩展的服务器日志。以上三种数据组建了数据提取、页面浏览、点击流和会话。页面浏览是指客户端用户一次点击网页的行为，一系列的页面浏览构成点击流。

网站结构和内容的处理是一个内部关联的任务。网页如何链接取决于网页的浏览方式，网站内容的创建技术又决定着网站的内容和结构，而不同的用户则决定着网站主页内容的设计。因此网站的结构、内容和用户的使用有着密不可分的联系，网站的结构和内容影响着 Web 使用挖掘的不同阶段，页面文件在语义上依赖着网站内容，而网站内容的决定是一个手工过程，取决于创建网站的技术和分析目的。

4　网络的使用挖掘

Web 使用挖掘预处理是将包含在不同应用数据源中的用户使用、网页内容和结构信息转换成适合于数据挖掘的形式。一旦这些用户使用数据被处理成会话文件，那么许多技术如：关联规则、相似页面或用户的聚类、序列模式等就都能使用了。反过来，挖掘出的模式又能被用于许多应用，比如：网站设计、个性化、安全检测和商业决策分析。模式分析是将数据挖掘结果转换成有用的和有趣的知识。

4.1　预处理

在数据的预处理过程中，数据清理是一个特定的处理阶段，包括从多服务器上合并日志数据，并且分析划分数据领域，通常在这个阶段剔除图文件的请求。另外用一些方法进行用户识别和会话识别，最常用的方法有 Cookies、用户注册和在 URIs 中插入会话 Ids。

页面浏览识别要确定哪些页面文件请求是同一个页面浏览的一部分或属于哪部分内容，以上的步骤都与网站的结构和内容紧密相关。最后一个步骤是路径完善，是指由于本地浏览器缓冲导致的页面引用的丢失，通过路径完善将那些丢失的重要的页面引用找回来。该步骤不同于其它的步骤的地方是将找回的数据加到日志中。每一步都是为了创建一个会话

（下转第128页）

Step2: if Supp(第i条规则前件) ≤ Supp(第i条规则后件) then

Step3:计算规则可信度

那么因为近一半的规则不用计算其可信度,更不用进行评价,所以不仅不会增加原挖掘一评价算法的负担,还在一定程度上减少了它的时空复杂性。此算法命名为 Prara2。

以小支持度的项集为前件的规则其可信度也比以它为后件的规则的可信度大,但是不能以哪条规则的可信度大为依据在对规则中决定取舍,因为可信度并不具备充分性因子的性质,即1.1节中的性质1,而且充分性因子是由可信度和规则后件支持度共同决定的。

结论 用相关性度量识别关联规则的相关性,已经被广泛地采用,但它不能识别作为规则前件和后件的两个项集中哪一个对另一个的出现具有更大的促进作用;可信度分析也只能说明对规则中哪一个的可信度高,仅此而已;孙海洪博士在他提出的 QAR_SQL 算法中采用了充分性因子进行评价,但也只是将相关性定量地给出了。本文利用相关性度量和充分性因子的关系和后者的性质提出了对规则取舍问题的解决方案,它的意义在于减少了领域专家的工作量,些微地推进了自动评价技术的发展,在一定程度上解决了 KDD 主流发展中存在的问题之一——领域专家的局限。这方面的工作还将继续进行。

参考文献

1 杨炳儒. 知识工程与知识发现. 冶金工业出版社,2000

2 Han Jiawei. Micheline Kamber. Data Mining: Concepts and Techniques. 高等教育出版社,2001

3 孙海洪. KDD 算法和启发型协调器的理论研究及其应用:[博士论文]. 北京科技大学,2001

4 王永庆. 人工智能原理与方法. 西安交通大学出版社,1998. 162~171

(上接第83页)

文件,但由于在数据收集中代理服务器和缓冲技术的影响,产生的会话文件无法做到很准确。

Episode 识别是一个有选择性的预处理步骤,一般是放在必处理步骤之后,是由 W3C 定义的一个用户会话的语义子集。

4.2 模式发现

一但用户会话和事务已被识别就可采用以下技术进行挖掘。

(1)*路径分析* 判断在一个 Web 站点中最频繁访问的路径,其它的相关路径可通过路径分析得出,利用这些信息还可以改进站点的结构设计。

(2)*关联规则和序列模式的发现* 使用关联规则发现相关性,利用相关性更好地组织站点内的 Web 空间,实行有效的市场战略。序列模式的发现:能够便于预测用户的访问模式,有助于开展这种模式的有针对性的服务。

(3)*分类和聚类* 分类规则可以识别一个特殊群体的公共属性的描述,并可以用来分类新的市场战略。聚类规则以概率分析为基础,发现客户访问网站的整体分布情况。

4.3 模式分析

模式分析的方法有:联机分析、可视化、知识查询和信息过滤。首先用模式分析工具将抽象的使用模式以直观、容易理解的方式展现给分析者,然后分析者利用知识查询语言,根据需要对挖掘过程加以限制,得到感兴趣的使用模式。比如限定某一领域进行挖掘,然后就这一领域挖掘出来的使用模式进行分析,得出感兴趣的结果。

信息过滤分两部分:objective 过滤和 subjective 过滤。objective 过滤处理用不同模式发现关联的数值型度量的变化,比如:支持度和兴趣度;subjective 过滤是用来处理使用挖掘通过分析网站内容和结构而形成访问网页的可信任度。对于 Web 使用挖掘,设想用网站结构和内容作为网站设计者的领域知识,在网页之间进行链接以提供这些页面的关联支持,那么在网页之间的拓扑链接越强,这些网页一起被访问的可信度也就越高。类似地,在同一个内容簇或同一类里的页面被认为在一起被访问的可信度远远大于不同簇或不同类中的页面。

结束语 本文进一步强调了 Web 的结构和内容不仅仅对使用挖掘有影响,而且是至关重要的和密不可分的。它是使用挖掘处理算法的重要数据源,贯穿整个使用模式发现的全过程,并且能为模式分析中的用户行为的预测提供信息。进一步讲,网站结构和内容的挖掘结果又可以作为很重要的数据源进行网页的分类和聚类,从而提高 Web 使用挖掘的效率。

参考文献

1 Cooley R ,Tan P-N, Srivastava J. Websift: The Web site information filter system. In WEBKDD,San Diego,CA,1999

2 Liu Lizhen. The Research of Web Mining. In: The 4th World Congress on Intelligent Control and Automation,2002

3 Cooley R,Mobasher B,Srivastava J. Data preparation for mining world wide Web browseing patterns. Knowledge and Information Systems,1999

4 Linoff G S,Berry M J A. Mining the Web,2001

5 Mena J. Data Mining Your Website,1999

6 Fayyad U,et al. The KDD process for extracting useful knowledge from volumes of data. Communications of the ACM, 1996, 39(11)

7 Hahn U, Schnattinger K. Deep knowledge discovery from natural language texts. In: Proc of the 3rd Int'l Conf. Knowledge Discovery and Data Mining ; Newport Beach, 1997

8 Wang Wei Qiang. Text Ming on the Internet Computer Science. 2000

9 Wang Ji Chen. Research on Web Text Mining. Journal of computer Research&Development, 2000,37(5)

10 Chen En Hong. Web Usage Mining: Discovering User Behavior Patterns From Web Data. Computer Science, 2001,28(5)

11 Yang Xiao Hua. Hyperlink Structure Mining of Web Sites. Computer Project&Application, 2001,8

12 Wang Shi. Web Mining. Computer Science, 2000,27(4)

13 Wang xiao ya. ,Web Usage Mining:[PH. D thesis]. 2000,3

14 Liu Jun. Web Usage Mining:[Master thesis]. 2000