基于免疫聚类的 Web 日志挖掘*

吕 佳
(重庆师范大学 数学与计算机科学学院,重庆 400047)

摘 要:Web 日志挖掘旨在使用数据挖掘技术从 Web 服务器日志文件中挖掘出有用的规律和模式,以此改进网站结构以及实现 Web 个性化服务。本文提出基于免疫聚类的 Web 日志挖掘算法,利用人工免疫系统的基本原理来进行用户聚类分析,从而发现相似客户群体、挖掘潜在客户。免疫聚类通过模拟免疫系统体液免疫应答的基本过程,提取出数据的基本特征,以此概括数据的分布特征,从而实现 Web 日志数据的无监督自组织聚类。通过在真实数据集上的实验证明了该算法的可行性和有效性。

关键词:Web 日志挖掘;数据预处理;用户会话;免疫系统;免疫聚类
中图分类号:TP391 **文献标识码**:A **文章编号**:1672-6693(2007)02-0032-04

Web Log Mining Based Upon Immune Clustering

Lü Jia
(College of Mathematics and Computer Science, Chongqing Normal University, Chongqing 400047, China)

Abstract: Web log mining aims at mining some useful rules and patterns from web server log file by means of data mining technology so that website structure can be improved and web personality service can be realized. An immune clustering algorithm of web log mining is presented in this paper. It applies artificial immune system to clustering analysis for the sake of finding similar customer population and mining underlying customer. Immune clustering simulates sap of immune response process of immune system and extracts the essential feature of data and generalize distribution feature of data, accordingly unsupervised clustering of web log mining can be realized. Experimental results applied on real data set show that immune clustering algorithm of web log mining is feasible and effective.

Key words: web log mining; data pre-processing; user session; immune system; immune clustering

随着 Internet 的迅速发展,Web 在人们的日常生活和工作中的地位日益显著。Web 中包含了 Web 页面的内容信息、超链接信息,以及 Web 页面的访问和使用信息,如何针对这些信息,应用数据挖掘技术挖掘出有用的信息,更好地为用户服务,已经成为目前国内外的一个新的研究热点[1]。按照 Web 挖掘对象的不同,分为 Web 内容挖掘、Web 结构挖掘、Web 日志挖掘。Web 日志挖掘的数据源主要是 Web 服务器日志文件,它是 Web 服务器用以记录用户访问该网站的各页面情况的文件,体现了用户使用 Web 的行为特点以及隐藏在用户行为背后的更深层次的动因和规律。通过对 Web 日志的挖掘,有助于更好地理解 Web 和 Web 用户访问模式,从而调整网站结构,有针对性地为用户推荐信息,提供个性化服务,这样对于开发 Web 的最大经济潜力都是非常关键的[2]。

人工免疫系统是模拟自然免疫系统功能的一种新的智能方法[3],提供噪声忍耐、无教师学习、自组织、自记忆的进化学习机理,其研究成果涉及到信息安全、模式识别、数据挖掘[4]、智能优化、控制和故障诊断等诸多领域,已经成为继神经网络、模糊逻辑和进化计算后的人工智能的又一研究热点。将人工免疫系统的基本原理引入到 Web 日志挖掘中,利用免疫系统的自学习和自记忆的特点,从而挖掘出 Web

* 收稿日期:2006-01-20
资助项目:重庆师范大学科研基金(No. 05XLY003)
作者简介:吕佳(1978-),女,四川达县人,讲师,硕士,研究方向为 Web 数据挖掘。

日志中有用的模式。

1 Web 日志挖掘过程分析

Web 日志挖掘过程一般分为 4 个阶段,即数据预处理阶段、日志挖掘算法实施阶段、模式分析阶段和可视化阶段[5],其流程如图 1 所示。

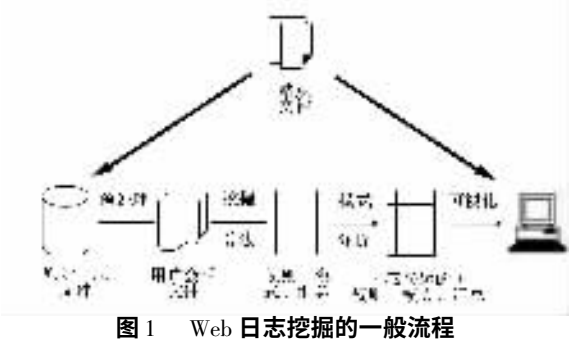

图 1 Web 日志挖掘的一般流程

数据预处理是将原始的日志文件转化为适合进行数据挖掘的可靠的精确的数据。日志挖掘算法实施阶段是对数据预处理的结果施用挖掘算法产生规则与模式。数据预处理和日志挖掘算法是 Web 日志挖掘中的关键技术,数据预处理的结果作为挖掘算法的输入直接影响日志挖掘算法产生的规则与模式。模式分析阶段分析挖掘得到的规则和模式,提取有意义的、感兴趣的规则与模式作为挖掘结果。可视化则是用图形化界面将挖掘的结果显示出来。

2 Web 日志挖掘中的数据预处理

W3C 组织规定了服务器日志的两种格式:通用日志格式(Common Log Format)和扩展型日志格式(Extended Log Format)。数据预处理就是要根据挖掘目的,对原始 Web 日志文件中的数据有针对性地进行提取、分解、合并,最后转化为适合进行数据挖掘的数据格式并保存到关系型数据库表或数据仓库中,等待进一步处理。这个环节是整个过程的基础和实施有效挖掘算法的前提,在 Web 日志挖掘中起着非常重要的作用。这一过程主要包括数据清洗、用户识别、会话识别、路径补充等步骤,如图 2 所示。

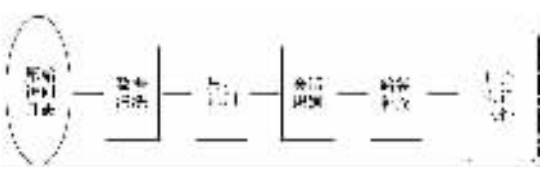

图 2 Web 日志挖掘中数据的预处理过程

2.1 数据清洗

当用户请求一个网页时,与这个网页有关的图片、音频等信息会自动下载并记录在日志文件中。数据清洗就是要删除这些无关紧要的嵌入式文件。一般采用的方法是删除与挖掘目的无关的数据,如删除后缀为.gif,.GIF,.jpeg,.JPEG,.jpg,.JPG,.map,.ICO,.ps,.SWF,.cgi,.js,.CSS 等各类嵌入式文件。此外,数据清洗阶段还要合并一些记录,对用户请求页面时发生错误的记录进行适当的处理等。日志中包含的状态(status)信息是由服务器记录的,以表示对一个请求的响应情况,200 一般表示成功,304 表示页面重定向,404 表示没有找到请求的文件即请求失败。故仅对成功请求到页面的日志项进行处理。

2.2 用户识别

由于本地缓存、代理服务器(包括网吧、局域网等环境)以及防火墙的使用,使得在日志中辨识每一个用户变得较为复杂。如不同的用户可以在同一时间通过一个简单的代理访问 Web 服务器;同一个用户可能在不同的机器上访问 Web 服务器;一个用户可能在同一台机器上使用不同的浏览器访问 Web 服务器;而不同的用户也可能使用同一台机器浏览某一站点。基于日志/站点方法是目前最为常用的 Web 日志挖掘工具采用的技术,并结合一些启发式规则来识别用户。只要 IP 地址、操作系统、浏览器、浏览器版本以及 http 协议版本中有一项不同即视为不同的用户。

2.3 会话识别

用户会话就是一个用户对 Web 服务器的一次有效访问,通过其连续请求的页面可以获得用户在网站中的访问行为和浏览兴趣。通常以时间长度来确定用户会话,实验中采用 30 min 来切分用户会话。一次会话中用户访问的最后一个页面停留时间均以 500 s 来计,同时若一个用户在某一页面上停留的时间不足 30 s,则认为该页面对分析用户行为没有贡献,将其删除。

2.4 路径补充

由于存在客户端缓存等原因,使得 Web 服务器日志中并没有完整地记录下用户的访问行为,路径补充就可以将遗漏的信息补充进来。例如,如果用户的请求是通过本地(如在 IE 浏览器中按“后退键”)或 Proxy 端的缓冲区得到满足,则服务器端无法记录此次请求,这样就需要进行路径补充,同时还可以结合站点的拓扑结构来补充路径。

表 1 为从 Web 日志中初步转化到数据库中的文件片断,经过完整的数据预处理过程得到用户会

话文件。再根据挖掘目的数据格式化为相应的可供　　日志挖掘算法实施挖掘的数据文件。

表 1　Web 日志中初步转化到数据库中的文件片断

	user_ID	IPaddress	URL	Time	browser	OS	protocol	delay
1	1	130.226.169.201	english/	6:35PM	Mozilla4.0	Windows NT 5.1	HTTP/1.1	500
2	2	158.182.1.30	english/	7:08PM	Mozilla4.0	Windows NT 5.1	HTTP/1.0	500
3	3	159.226.231.68	main023.htm	7:27PM	Mozilla4.0	Windows NT 5.1	HTTP/1.1	500
4	4	172.30.11.31	main013.htm	6:33PM	Mozilla4.0	Windows NT 5.1	HTTP/1.1	500
5	5	172.30.12.13	guanlijigou/erpsoxuanpeibao/index.htm	6:46PM	Mozilla4.0	Windows NT 5.1	HTTP/1.1	300
6	5	172.30.12.13	guanlijigou/kejiqiye/english/sindex.htm	6:46PM	Mozilla4.0	Windows NT 5.1	HTTP/1.1	23
7	5	172.30.12.13	guanlijigou/kejiqiye/images/b8.png	6:46PM	Mozilla4.0	Windows NT 5.1	HTTP/1.1	0
8	5	172.30.12.13	guanlijigou/kejiqiye/index.htm	6:46PM	Mozilla4.0	Windows NT 5.1	HTTP/1.1	6
9	5	172.30.12.13	guanlijigou/kejiqiye/keyuan/index.htm	6:46PM	Mozilla4.0	Windows NT 5.1	HTTP/1.1	31
10	5	172.30.12.13	guanlijigou/xiaoyiyuan/default.htm	6:24PM	Mozilla4.0	Windows NT 5.1	HTTP/1.1	0

3　基于免疫聚类的 Web 日志挖掘

3.1 免疫系统基本原理[6]

免疫系统是抗击病原体入侵的首要防御系统,包括许多补体种类的免疫细胞及制造这些免疫细胞的免疫器官,为数最多的免疫细胞是淋巴细胞。淋巴细胞主要包括 B 细胞和 T 细胞。能被 B 细胞和 T 细胞识别并刺激二者进行特异性应答的病原体,叫做抗原。T 细胞识别特异抗原后,一方面 T 细胞复制并激活杀伤 T 细胞,杀伤 T 细胞杀死任何被特异抗原感染的细胞,这称为细胞免疫;另一方面通过辅助 T 细胞激活 B 细胞,激活后的 B 细胞识别特异抗原,并克隆扩增分化为浆细胞产生抗体,抗体与抗原结合,杀死抗原,这称为体液免疫。

体液免疫应答反映了免疫系统中抗体与抗原、抗体与抗体之间的相互作用关系和抗体学习抗原的行为特性。这种免疫应答过程主要包含了克隆选择、细胞克隆、记忆细胞获取、亲和力突变、克隆抑制及动态平衡维持等机制。这些机制相互作用来实现免疫系统的防御能力。

3.2　基于免疫聚类的 Web 日志挖掘

免疫聚类是借鉴生物免疫系统中体液免疫应答过程中的基本免疫机制的一种自组织聚类算法[7]。利用免疫系统的自组织、自学习和自记忆的特性,来学习数据的分布特性,从中选择或产生能概括分布特征的较少的样本。

为用户会话中的每一个 URL 分配一个唯一的标识,$id \in (1,2,3,\cdots,u)$,其中 u 是所有有效的 URL 的总和。这样用户会话可以用具有 u 位二进制串来表示。第 i 个用户会话二进制位串的表示方法为

$$S_{if}=\begin{cases}1 & \text{在第 } i \text{ 个会话中用户访问了第 } j \text{ 个 URL}\\ 0 & \text{在第 } i \text{ 个会话中用户没有访问 第 } j \text{ 个 URL}\end{cases} \quad (1)$$

计算两个用户会话亲和力的方法采用(2)式[8]

$$S_{kl}=\frac{\sum_{i=1}^{u}S_{ki}S_{li}}{\sqrt{\sum_{i=1}^{u}S_{ki}}\sqrt{\sum_{i=1}^{u}S_{li}}} \quad (2)$$

数据预处理后形成的用户会话集即为抗原 Ag。这样,免疫聚类的算法流程如下。

step1 初始化操作。随机生成群体规模为 N 个 u 位的二进制形式的初始抗体集 Na,确定亲和力阈值 $Q1$,网络抑制阈值 $Q2$,克隆选择率 a,其中 $a \in (0,1)$,迭代次数 T。初始化总体记忆抗体网络集 Abm 为空集 ϕ。

step2 计算亲和力。对 Ag 中任一抗原 Agi,按照(2)式计算 Na 中每个抗体和抗原 Agi 的亲和力,并按照亲和力大小排序。

step3 克隆选择。从 Na 中选择亲和力最高的 $a*N$ 个抗体构成克隆源。依据亲和力成正比的关系进行克隆操作,克隆数为亲和力的增函数,得到 Nb 个抗体克隆集。

step4 高频变异。对抗体克隆集 Nb 中的抗体以反比于亲和力的概率进行变异操作,得到 Nc。

step5 再次计算亲和力。计算 Nc 中的抗体对应于抗原 Agi 的亲和力,按照亲和力大小排序,选择亲和力最高的前 $\eta\%$ 且亲和力阈值大于 $Q1$ 的抗体形成免疫记忆参考集 Ad。

step6 克隆抑制。计算 Ad 中抗体之间的亲和力,删除亲和力大于 $Q2$ 的一方抗体,另一方保留。合并到记忆抗体集 Abm。

step7 选择下一个抗原,转到 step2,直到所有抗原均计算完毕,然后再转到 step8。

step8 按照 step6 的方法对 Abm 进行克隆抑制。

Step9 若满足终止条件,则输出记忆抗体集 Abm,结束算法。否则随机生成 d 个抗体,将其合并到 Abm 中作为新一代抗体集 Na,转 step2 继续迭代。

4 Web日志挖掘实验结果

为了验证本算法的有效性,实验对象采用重庆大学校园网2004年12月6日的部分数据。原始日志记录有44 717条,经过数据清洗后有1 562条。最后识别出25个用户会话和124个页面。表2为实验结果。

表2 Web日志挖掘实验结果

免疫聚类	聚类数	迭代次数	准确率
	8	50	80.3%

从表中实验结果来看,基本上可以聚类出具有相同访问兴趣度的用户,如其中有两大类用户分别集中在对院系设置和机构管理的访问上。与一般的聚类算法不同,免疫聚类算法挖掘相关页面集不需要精确地确定一个用户会话属于哪一个聚类,而是通过提取数据的分布特征,尽可能得到近似聚类中心。用这些能代表一个类的较少的几个典型的抗体来概括出一个类的特征。为下一步进行模式分析提供依据。

5 结束语

Web服务器日志文件中包含了用户浏览信息,通过对Web日志进行各种定量或定性的分析,揭示隐藏在这些信息背后的各种关系,从而帮助站点的管理者或经营者制定相应的策略,向用户提供个性化服务,提高Web服务质量[9]。聚类分析作为数据挖掘中的一项重要技术,适合于发现相似客户群体、挖掘潜在客户。由生物免疫系统启发而来的人工免疫系统具有很强的自学习、识别、自记忆和特征提取能力,适合于进行数据挖掘。由此,本文提出基于免疫聚类的Web日志挖掘算法,通过模拟免疫系统体液免疫应答的基本过程从而实现Web日志数据的无监督自组织聚类。通过在真实的数据集上的实验已经证明了本算法的有效性。而本算法和其它相关聚类算法在效率和准确性上的对比实验及其对本算法的改进是下一步将要开展的工作。

参考文献:

[1] 郭岩,白硕,于满泉. Web使用信息挖掘综述[J]. 计算机科学,2005,32(1):1-7.

[2] 宋擒豹,沈钧毅. Web日志的高效多能挖掘算法[J]. 计算机研究与发展,2001,38(3):328-333.

[3] 肖人彬,王磊. 人工免疫系统:原理、模型、分析及展望[J]. 计算机学报,2002,25(12):1281-1293.

[4] CASTRO L N de, VON ZUBEN F J. An Evolutionary Immune Network for Data Clustering[A]. JANEIRO P de. Proc of the 6 'Brazilian Symposium on Neural Network[C]. Brazil, 2000. 84-89.

[5] 费爱国,王新辉. 一种基于Web日志文件的信息挖掘算法[J]. 计算机应用,2004,24(6):57-59.

[6] 李涛著. 计算机免疫学[M]. 北京:电子工业出版社,2004.

[7] CASTRO L N de, VON ZUBEN F J. Learning and Optimization Using the Clonal Selection Principle[J]. IEEE Trans on Evolutionary Computation, Special Issue on Artificial Immune Systems, 2002,6(3):239-251.

[8] 恽爽,韩立新,董浚,等. KDW综述:基于Web的数据挖掘[J]. 计算机工程,2003,29(1):284-286.

[9] 吕佳. Web日志挖掘技术应用研究[J]. 重庆师范大学学报(自然科学版), 2006,23(4):43-45.

(责任编辑 李若溪)