Web日志挖掘中数据预处理技术的研究

于 飞,丁华福,姜 伦
(哈尔滨理工大学 计算机科学与技术学院,黑龙江 哈尔滨 150080)

摘 要:数据预处理在Web日志挖掘过程中起着至关重要的作用,直接影响日志挖掘的质量和结果。详细分析了数据预处理的过程,提出一种改进的数据清洗方法,以提高日志挖掘中数据预处理的效率,并针对Web日志数据预处理中会话识别这一重要环节,提出一种改进的会话识别方法。在用户识别后,根据页面内容、站点结构确定页面重要程度,对阈值进行调整。然后,根据用户对页面内容的兴趣度来删除会话中的链接页面和不感兴趣的页面。实验结果表明,提出的方法能更准确地确定页面访问时间阈值,得到更为合理有效的会话集合。

关键词:Web日志挖掘;数据预处理;会话识别;数据清洗
中图分类号:TP311 **文献标识码**:A **文章编号**:1673-629X(2010)05-0047-04

Research on Data Preprocessing Technology in Web Log Mining

YU Fei, DING Hua-fu, JIANG Lun
(College of Computer Sci. & Tech., Harbin Univ. of Sci. & Tech., Harbin 150080, China)

Abstract: Data preprocessing plays an essential role in the process of Web log mining, directly influenced the quality of the Web log mining and its results. Analyses data preprocessing process for Web log mining in detail, proposes an improved method of data cleaning, to improve the efficiency in data preprocessing of log mining, and proposes an improved method of session identification to Web log data preprocessing. The threshold is adjusted by the page weightness based on site's structure after the user identification. Then, delete the link pages and uninterested pages based on the user's interest degree of page content. Experimentally, the approach proposed can decide the access time threshold more accurately. It is more reasonable and effective.

Key words: Web log mining; data preprocessing; session identification; data cleaning

0 引 言

随着Internet信息量的剧增,如何帮助用户快速有效地获取自己感兴趣的信息,已成为网站设计者亟待解决的问题。解决这个问题的途径之一就是将数据挖掘技术和Web结合起来,进行Web挖掘。作为Web挖掘的一个重要组成部分,Web日志挖掘就是通过分析用户访问Web时在服务器留下的访问记录来发现用户访问Web页面的模式,帮助用户在海量的信息中寻找感兴趣的内容,实现"信息找人,按需服务"的个性化推荐。对于门户网站、电子商务类网站来说,可以更好地发现用户的兴趣所在,提高网站的服务质量,提高用户的忠诚度,从而提高网站的核心竞争力。

Web日志挖掘过程主要分为3个阶段:数据预处理、模式发现、模式分析[1]。数据预处理的目的就是将原始日志记录经过处理形成用户的会话文件,为模式发现算法实施阶段作好数据准备。会话识别是数据预处理中最重要的环节,会话识别的准确与否直接影响了后续工作是否能得到理想的结果,同时也决定了最终挖掘出的知识的可信度。文中提出了一种新的会话识别方法,该方法基于页面内容和站点结构,通过对页面的链入、链出数等几个参数的综合,得到每个用户页面的访问时间阈值,根据该阈值来切分用户会话,得到会话候选集合;然后,根据用户对页面内容的兴趣度来删除会话中的链接页面和不感兴趣的页面,生成一种最终有效的访问页面序列,从而为以后的模式发现提供良好的数据。

1 Web日志数据预处理过程

数据预处理是在将Web日志文件转换成数据库文件以后进行的,其目的是把Web日志转化为适合进行数据挖掘的可靠的、精确的数据。这个过程主要包括5个阶段:数据清洗、用户识别、会话识别、路径补

收稿日期:2009-08-31;**修回日期**:2009-11-12
基金项目:国家自然科学基金项目(60736014)
作者简介:于 飞(1983-),男,黑龙江哈尔滨人,硕士研究生,研究方向为数据挖掘;丁华福,硕士生导师,教授,研究方向为数据库、数据挖掘。

充和事务识别[2]。数据预处理过程见图 1。

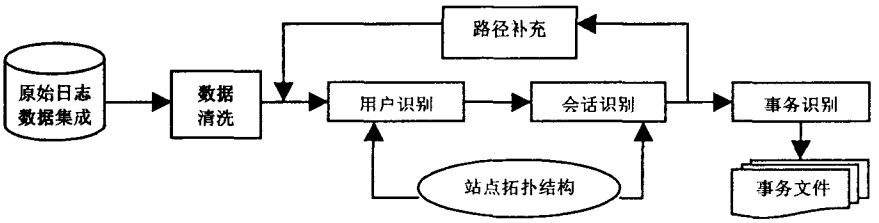

图 1 数据预处理过程

1.1 数据清洗

数据清洗是整个 Web 日志预处理的基础,主要是将有噪声的、不一致的、无关的数据从 Web 日志数据源中清除,合并某些记录,然后进行分析并将它们存入相应的数据字段中,删除 Web 服务器日志中与挖掘算法无关的项[3]。文中提出以纵向缩减和横向缩减对原始日志数据进行清洗,以提高日志挖掘中数据预处理的效率。

1.1.1 纵向缩减

根据以下三个方面对日志数据进行纵向缩减(行缩减):

1)后缀过滤:一般信息网站中,只是 html 文件与用户会话有关,后缀名为 gif,jpeg,jpg 的图片文件,后缀名为 wav 的音频文件以及后缀名为 cgi,js 的脚本文件对后面的分析处理不造成任何影响,将其删除。如果是一些特殊的网站如图片网站,可以重新设置相关的信息,若要分析网站流量,则这些记录都必须保留。

可根据网站的分类,在规则库中定义相应的清洗规则,或者修改已有的清洗规则,从而使其适用于不同的数据源清洗,使之具有较强的通用性和适应性。

2)方法过滤:GET 方法是用户请求页面的操作,其它的操作如 POST 等则可以过滤掉。

3)状态码过滤:状态码指示用户请求的结果。以 2 开头的表示请求成功,如 200 表示交易成功;以 3 开头的表示请求被成功转向,如 302 表示找到请求的页面;以 4 开头的表示链接出错,如 400 表示错误请求;以 5 开头的表示产生服务器错误,如 500 表示服务器产生内部错误。在进行数据清洗时应该过滤掉以 4 和 5 开头的信息。

1.1.2 横向缩减

在对日志数据进行清理时,通常只考虑到使用纵向缩减来减少日志文件的大小,但在使用数据挖掘算法进行挖掘任务时并不需要 Web 日志记录的众多属性。如要使用挖掘算法对 Web 站点的流量进行分析时,只需要保留用户访问时间、用户请求的 URL 页面等属性;而对用户进行聚类分析时,只需要保留用户访问时间、用户 IP 地址、用户请求访问的 URL 地址以及用户所使用的代理。与纵向缩减相对应,将这种缩减日志记录中属性的方法定义为横向缩减(列缩减)。横向缩减只会减少日志记录数据表中属性列,不会缩减日志记录的行数。

纵向缩减为数据预处理的以下步骤打下基础,而横向缩减则为预处理的以下步骤提供方便。数据表 log080615 大小为 1652k,具有 11685 条日志记录,在未实施数据清洗前,该数据表中包含请求页面文件记录、请求站点页面上图形文件记录以及用户请求失败的记录。根据以上方法对该表实施纵向缩减后数据表 log080615 大小为 226k,具有 1236 条日志记录。若在对 Web 日志进行分析的主要目的是对用户进行聚类分析,则再对数据表实施横向缩减,保留用户访问时间、用户 IP 地址、用户请求访问的 URL 地址以及用户所使用的代理属性,则缩减后数据表 log080615 大小为 168k,具有日志记录 1236 条。

实验证明从纵横两个方向对日志记录进行清洗不会降低日志清洗的精度,因为横向缩减只作用于日志记录的属性列上,但可以减少日志文件的大小,这样在应用挖掘算法时就可以减少算法扫描日志的时间,提高挖掘效率。

1.2 用户识别

用户识别的主要任务是消除代理或防火墙的影响,从日志中还原出每一个请求的用户。在对用户进行访问模式挖掘或用户聚类分析时,用户识别显得至关重要,因为群体是由个体组成的,只有对个体有了清楚的了解,才能识别群体的特征[4]。

可以通过下列启发式规则来识别用户:

1)判断用户 IP 地址,不同的 IP 地址代表不同的用户;

2)若 IP 地址相同,但是用户的浏览器或操作系统不同,则认为是不同的用户;

3)若 IP 地址相同,用户的浏览器和操作系统也相同,则根据引用页判别;如果用户请求的某个页面未被请求过且该记录中引用页为空,则认为是新用户。

但是,以上规则不能保证准确地识别出每一个用户。例如一个用户使用多种浏览器,通过几个代理服务器或者几台不同的机器访问 Web 服务器,将被认为是不用的用户。具有相同 IP 地址的几个用户使用同种操作系统和同种浏览器,并且看过的页面集合也一样,则会被认为是用一个用户。所以说,要准确识别出用户是一件非常困难的工作,其结果直接影响到整个

预处理结果。

1.3 会话识别

一个会话就是用户从进入到离开站点的一系列浏览请求。在跨越时间段较大的 Web 服务器日志中,用户可能多次访问了该站点,会话识别的任务就是把属于同一用户的同一次访问请求识别出来,即将用户的访问记录分为单个的会话[5]。

会话的划分方法很多,有些依据时间,有的依据站点拓扑结构。现有的启发式会话识别方法分为三类[6],前 2 个基于时间,后一个基于网站结构:

(1)给用户一个页面的请求时间规定一个阈值 δ,如果两个请求时间超过这个阈值则认为是新的会话的开始。设 t' 为当前会话最后一个请求的时间戳,如果下一请求的时间戳 t'' 满足 $t''-t'\leqslant\delta$,则加入当前会话,否则成为另一个会话的开始。通常 δ 取 10min。

(2) 给用户一次会话时间规定一个阈值 θ,如果超过这个阈值,则认为新会话的开始。设 t_0 为会话初始页的时间戳,一个 URL 请求的时间戳 t 如果满足 $t-t_0\leqslant\theta$,则该请求将被加入到本次会话,第一个满足 $t-t_0>\theta$ 的页面成为下一会话的初始页。实际应用时一般将 30min 作为缺省的阈值,而 J. PitKow 等做试验所得数据指出 25.5min 更为合适。

(3) 基于日志请求的参考栏,给连续页面访问时间间隔一个上界 Δ,设 p、q 为两个连续的请求,p 属于会话 S,t_p、t_q 为 p、q 的时间戳($t_p<t_q$),如果 q 的引用页在 S 中或 $t_p-t_q<\Delta$,则加入 S,否则 q 为另一会话的开始页。

2 会话识别的改进

考虑到不同的用户的行为差异,如网络位置、上网习惯、阅读速度、使用计算机的熟练程度等一系列因素,每个用户的会话时间通常是不一样的。例如,不同的用户浏览完页面的内容的时间是不一样的,同一个用户在不同的时间段对页面内容的兴趣度不一样,因此他们在页面的停留时间也会不一样。而在传统的基于时间的会话识别方法中,一般判断访问记录间的时间阈值是固定的,没有充分考虑到用户的个体差异。这样对于长会话中的超过时间阈值的请求页面会被划分到下一个会话中。

由于页面的停留时间显然与页面的内容和页面的结构有关,文中提出了一种新的会话识别方法。该方法基于页面内容和站点结构,通过对页面的链入、链出数等几个参数的综合,得到每个用户页面的访问时间阈值,根据该阈值来切分用户会话,得到会话候选集合;然后,根据用户对页面内容的兴趣度来删除会话中的链接页面和不感兴趣的页面,生成一种最终有效的访问页面序列。

2.1 结合页面内容与站点结构

页面的站点结构通常使用页面的链入、链出数来衡量页面重要性。页面的链入数指链接到某页面的页面个数,链出数指通过该页面到达其他页面的个数。

定义 1:链接内容比 LCR(Link-Content Ratio)是指页面链入链出数与页面内容之比。链入数是指链接到某页面的页面个数,记为 L_I,链出数是指某页面所包含的链接个数,记为 L_O,页面大小记为 PS(Page Size),则 LCR 计算公式为:

$$LCR=(L_I+L_O)/PS \tag{1}$$

一般情况下,一个页面的链入要比链出重要,因此需要对它们进行加权调整。文中以黄金分割来假设链入与链出的权值之比,式(1) 调整为:

$$LCR=2(0.618L_I+0.382L_O)/PS \tag{2}$$

为了将 LCR 值用于对阈值 δ 的调整,需要将 LCR 值映射到(0,1) 之间。可以选用多种映射方式,例如用 LCR 值与所有 LCR 值中的最大值的比值即可映射到(0,1) 之间,但这种方法容易受到孤立点的影响,当某个页面的 LCR 值很大时,会影响到其他点。文中选择如下的映射方式,记 LCR 对 δ 的影响因子为 β。

定义 2:β 为页面 LCR 值对页面访问时间阈值 δ 的影响因子,其计算公式为:

$$\beta=1-\exp(-\mathrm{sqrt}(\mathrm{sqrt}(R_{LCR}))) \tag{3}$$

2.2 阈值的确定和删除无关页面

2.2.1 阈值的确定

由于用户访问网页的时间为下载时间与正常阅读时间之和,即 T_d+T_r,但是对于连接速度比较慢的客户端或者页面产生时间较长的请求,用户可能在页面没有下载完成之前就开始阅读,因此使用 T_d 作为用户实际开始阅读时间会对这类会话的识别产生一定的误差。这就需要对 T_d 进行平滑处理,即使用 αT_d 作为用户开始阅读时间,其中 α 为平滑系数[7],实验时选择 0.2 ~ 1.0 作为平滑系数分别进行计算。实验表明,α 的值选取 0.72 比较合理。因此,可得阈值计算公式为:

$$\delta=(\alpha T_d+T_r)(1+\beta) \tag{4}$$

2.2.2 删除无关页面

定义 3:(浏览兴趣度,记作 P) 设定 P_j 表示用户在页面 j 上的浏览兴趣度,Count_{ij} 表示用户从页面 i 进入页面 j 的浏览次数,Time_{ij} 表示用户从页面 i 进入页面 j 的浏览时间(单位:秒),Sb_{ij} 表示从页面 i 进入页面 j 接收到的字节数(单位:MB),则用户浏览兴趣度:

$$P_j=\frac{\mathrm{Count}_{ij}\times\mathrm{Time}_{ij}}{\mathrm{Sb}_{ij}} \tag{5}$$

用户浏览网站时对某一页面感兴趣程度通常由页面的浏览时间、浏览次数和网速等几方面的因素决定。定义中的浏览时间在 Web 日志中指页面的耗用时间，页面的浏览时间越长说明用户对该页面越感兴趣，而浏览时间又与浏览速度有关，页面的浏览速度在 Web 日志中则对应页面的接收字节数。这个浏览兴趣度的定义更能全面地反应出用户对页面的关注程度。用兴趣度 P 进行比较删除会话中用户兴趣度不高的页面和链接页面。

2.3 识别方法的算法描述

要动态设置每个页面的访问时间阈值 δ，需要通过定义 1 获得该页面的正常阅读时间 T_r，结合页面的链接情况 LCR 的影响因子 β 调整 δ，其算法步骤如下：

(1) 根据 1.1 节提出的数据清洗方法，以纵向缩减和横向缩减对原始日志数据进行清洗；

(2) 根据日志文件进行用户识别；

(3) 针对每一个用户访问特定页面的记录，从日志文件中得到该用户的下载时间 T_d；

(4) 根据定义 2 的方法得到页面的影响因子 β，将 β、T_d 和 T_r 的值利用上述阈值的计算方法得到每个页面访问时间阈值 δ；

(5) 根据阈值划分日志文件，得到用户会话集合；

(6) 根据 2.2.2 节用户对页面的兴趣度 P 的比较，删除会话中用户兴趣度不高的页面和链接页面。

3 实验结果与分析

本节将对给出的改进的会话识别算法进行实验并加以分析。实验中的日志文件是黑龙江某高校的外网用户访问日志。本实验比较了这 4 种会话识别算法：基于固定的先验阈值的会话识别算法，基于页面访问时间阈值 δ 的会话识别算法，基于引用会话识别算法，以及基于页面与站点结构动态设置时间阈值 δ 的会话识别算法。

文中使用目前常用的评价标准[8]：会话被算法 h 完整重建的程度。一般使用精确度和查全度这两个指标衡量重建程度。精确度用完全构建的会话数目与构造生成的总会话 Rh 数目的比值表示：precision(h) = |Rh∩R|/|Rh|，查全度用完全构建会话数目与真正的会话 R 数目的比值表示：recall(h) = |Rh∩R|/|R|。实验数据如表 1 所示，在对各种算法比较时，以基于引用的方法为基准。

从表 1 的实验数据可以看出，改进后的方法无论从精确度还是从查全度来说，占的比例都最大，更能对日志进行合理的划分，有效地提高了会话识别的质量，而且更接近用户的真实会话。

4 结束语

文中提出的数据清洗方法，可以减少算法扫描日志的时间，提高挖掘效率；改进的会话识别方法，通过下载时间、阅读时间等多个参数来综合定义用户对页面的访问时间阈值，从而可以更加个性化地识别出会话，能够更准确地反映出用户对页面的关注程度，能够识别出大部分正常的长会话和非正常的短会话。同时，依据用户对页面的浏览兴趣度来删除会话中的无关的页面和链接页面，提高了下一步数据挖掘的效率和质量。

表 1 实验数据

会话构造方法	有效页面数	会话数	会话交集数	精确度(%)	查全度(%)
基于引用	126268	|R| = 24185	|R∩R| = 24185	|R∩R|/|R| = 100	|R∩R|/|R| = 100
基于固定时间阈值(T)	126268	|T| = 41238	|T∩R| = 11230	|T∩R|/|T| = 27.23	|T∩R|/|R| = 46.43
基于页面访问时间阈值(A)	126268	|A| = 56325	|A∩R| = 18583	|A∩R|/|A| = 32.99	|A∩R|/|R| = 76.84
改进后的会话识别方法(C)	96634	|C| = 51261	|C∩R| = 19362	|C∩R|/|C| = 37.77	|C∩R|/|R| = 80.06

参考文献：

[1] Srivastava J, Cooley R, Deshpande M, et al. Web Usage Mining : Discovery and Applications of Usage Patterns from Web Data[J]. SIGKDD Explorations, 2000, 1(2): 12 - 23.

[2] 王书舟，高中文. Web 使用挖掘技术在电子商务中的应用[J]. 微机发展(现更名：计算机技术与发展)，2003，13(2)：41 - 43.

[3] 赵 伟，何丕廉，陈 霞，等. Web 日志挖掘中的数据预处理技术研究[J]. 计算机应用，2003，23(5)：62 - 67.

[4] 熊忠阳，周亚峰. Web 访问挖掘的预处理技术的研究[J]. 计算机技术与发展，2007，17(8)：11 - 14.

[5] Facca M, Lanzi P L. Mining interesting knowledge from weblogs : A survey[J]. Data and Knowledge Engineering, 2005, 53(3): 225 - 241.

[6] He D, Goker A. Detecting session boundaries from Web user logs[C]// Proceedings of the 22nd Annual Colloquium of IR Research. Cambridge: [s. n.], 2005: 57 - 66.

[7] 殷贤亮，张 为. Web 使用挖掘中的一种改进的会话识别方法[J]. 华中科技大学学报：自然科学版，2006，34(7)：33 - 35.

[8] Cooley R, Mobasher B, Srivastava J. Data Preparation for Mining World Wide Web Browsing Patterns[J]. Knowledge and Information System, 1999, 1(1): 32 - 40.