文章编号: 1000-5862(2012)04-0395-04

改进的页面与时间阈值的会话识别法

郑立山, 滕少华*

(广东工业大学计算机学院, 广东 广州 510006)

摘要: 在常用的计算时间阈值识别方法的基础上, 提出了一种改进的基于 URL 页面类型、页面信息量和访问时间的平均阈值识别方法. 针对不同的页面类型采用不同的阈值计算方法设置时间阈值, 相对于已有的对所有用户页面使用单一的先验阈值和现有动态阈值计算, 该方法能够更真实地反映用户会话的情况, 且识别的准确率有了较大提高.

关键词: Web 日志挖掘; 数据预处理; 用户会话识别; 动态阈值
中图分类号: TP 301 文献标志码: A

0 引言

随着电子商务和SNS网络社区的发展, 网站提供的信息越来越丰富多样, 对用户访问日志信息的挖掘显得更加突出. 在这类挖掘中, 会话识别是重要的预处理步骤, 处理不当而引入不良数据将直接影响到后续的挖掘工作.

本文主要研究会话识别的方法, 相关工作如下:

(1) 文献[1]提出了基于统计特征的会话识别方法, 通过使用统计值来动态设定访问时间阈值以识别会话.

(2) 文献[2]提出了通过过滤框架页面大幅度减少实验产生的有效页面数, 为每个页面设置访问时间闭值, 并根据页面重要程度对该阈值进行调整, 页面的重要性由页面内容及站点结构确定.

(3) 文献[3]提出了一种改进的基于时间间隔的识别方法. 该方法通过使用访问时间间隔超出某个阈值来识别会话, 该方法利用页面访问时间的正态分布来计算阈值, 比单一的先验阈值有所改进, 但当访问的站点有较多导航目录时, 如电子商务类型的网站, 该算法存在较大误差.

(4) 文献[4]提出一种基于待挖掘站点首页的用户会话识别方法, 该方法以站点首页作为用户新会话开始的标志, 但缺乏一定的合理性, 由于很多站点的起始页不一定是首页, 可能是通过搜索引擎等其他站点过来的.

本文主要研究基于页面媒体类型时间阈值的会话识别方法. 现有的基于时间阈值的会话算法有动态的, 也有固定的. 动态阈值会话识别的效率要高于固定阈值效率, 但是现有动态阈值的会话识别算法仍存在以下问题: (1) 动态阈值计算方法单一, 而用户访问的资源是多样化的; (2) 现有的动态阈值计算很难适应现有的网站应用, 如SNS社区网站、视频网站、电子商务网站等.

基于页面媒体类型时间阈值的会话识别法是一种针对目前主流网站, 如电子商务、SNS、门户网站等具有清晰页面结构的会话识别算法, 但实际页面中既有文字型, 也有多媒体型和超链接型的页面, 因而用统一文字型处理方法, 难以反映用户的实际操作.

据此, 本文将URL定义为以下3种形式: 导航型URL、文字型URL和多媒体型URL, 并根据不同的页面特点, 分别采用不同的时间阈值来计算不同类型URL页面, 实验验证了本文的方法能更好地反映用户操作.

1 改进的页面与时间阈值的会话识别法

1.1 概念与符号描述

在本文中, 将访问URL划分成导航型URL、文字型URL和多媒体型URL等3种类型. 针对不同的URL类型, 分别采用不同的页面时间阈值计算方法,

收稿日期: 2012-03-09
基金项目: 广东省自然科学基金(06021484, 9151009001000007)和教育部重点实验室开放基金(110411)资助项目.
作者简介: 滕少华(1964-), 男, 江西省南昌人, 教授, 博士, 主要从事数据挖掘和网络安全的研究.

以便能客观地反映用户实际操作.

(1) 导航型URL: 导航型URL是指链接到导航页面的URL链接, 该页面为其他页面提供链入地址, 以方便用户浏览, 如电子商务网站的分类页面. 该类型页面的特点是提供链接入口, 并不提供详细的信息. 对于这类页面, 用户更关注里面提供的链接的价值. 针对该类型的URL, 本文更注重页面内容与站点结构对页面阈值 δ 的影响.

(2) 文字型URL: 该页面用于展示详细信息, 其内容以文字为主, 如新闻页面、博客页面等. 对于这类页面, 一般用户停留的时间较长, 用户更关注文本的信息. 针对该类型的URL, 本文考虑页面内容与页面停留时间对页面阈值 δ 的影响.

(3) 多媒体型URL: 该类型的URL指向的页面主要以图片、视频、声音、网页游戏等多媒体信息为主, 该页面的特点是内容丰富, 形式多样, 用户在这类页面上停留的时间很长.

在生成时间阈值时, 根据页面的特点采用不同的阈值计算算法, 如表 1 所示.

表 **1** 页面类型-算法

页面类型	算法
导航型 URL	基于页面内容与站点结构的阈值计算法
文字型 URL	基于页面内容与页面停留时间的阈值计算法
多媒体型 URL	基于页面停留时间的阈值计算法

通过以上分类处理能更有效地模拟用户的真实会话情景. 在处理文字型 URL 时提出页面内容与页面停留时间阈值计算法, 该方法考虑页面的重要性与用户的访问时间和页面的内容大小有关. 如果页面大小相同, 其访问时间越长, 页面内容越重要; 如果访问时间相同, 页面内容越少, 页面越重要.

通过检验证明, 页面的访问时间 t 呈正态分布.在这个基础上, 选择能覆盖 94%样本集的时间数据值 t 作为参考值, 再乘以一定的平滑系数 α 后得到 δ. 选择 94%这个值是为了尽量不考虑 t 中的异常值, 选择一定的平滑系数是为可能删除一些合理的数据 t 而做的一种折中处理. 实验表明, 选择 α 为 1.0~1.5 作为平滑系数较好, 本实验中选择 α 为 1.25.

平滑系数和页面访问时间 t 的生成主要用于阈值 δ 的计算.

1.2 基于页面内容与站点结构阈值计算法

定义**1** 链接内容比(R_{LCR})[2]是指页面链入链出数与页面内容之比, 记页面大小为 S_{DS}, 则 R_{LCR} 计算公式为 $R_{LCR}=(L_I+L_O)/S_{DS}$.

这里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. 若一个页面的大小 S_{DS} 为3 kB, 它的链入数 L_I 为5, 链出数 L_O 为4, 则 R_{LCR} 计算结果为3. 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. 本文以黄金分割来假设链入与链出的权值之比, R_{LCR} 计算公式调整为 R_{LCR} =2(0.618 L_I + 0.382Lo)/ S_{DS} .

为了将 R_{LCR} 值用于对阈值 δ 的调整, 需要将 R_{LCR} 值映射到(0, 1)之间. 可以选用多种映射方式, 例如用 R_{LCR} 值与所有 R_{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 R_{LCR} 值很大时, 会影响到其他点. 本文选择如下的映射方式, 记 R_{LCR} 对 δ 的影响因子为 β.

定义**2** β 为页面链接内容比 R_{LCR} 对页面访问时间阈值的影响因子[2], 其计算公式为

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

定义**3** δ 为页面访问时间阀值[12], 其计算公式为 $\delta=\alpha\, t(1+\beta)$, 其中 α 为平滑系数, t为页面访问时间, β 为影响因子.

1.3 基于页面内容与页面停留时间阈值计算法

定义**4** 时间内容比(R_{TCR})是指用户在该页面的平均停留时间与页面内容(S_{DS})之比, 则 R_{TCR} 的计算公式为 $R_{TCR}=\sum_{i=0}^{n} t_i / n \Big/ S_{DS}$.

这里将第i个用户在某页面的停留时间记为t_i. 若一个页面在4个可识别的会话中的停留时间分别为3、5、4; 一个页面的大小为3 kB, 则 R_{TCR} 的值为 4.

为了将 R_{TCR} 用于对阈值 δ 的调整, 同样将需要将 R_{TCR} 值映射到(0, 1)之间, 可以选用多种映射方式, 因为 R_{TCR} 取的是有效会话, 已经排除了孤立点对它的影响, 采用如下方式将 R_{TCR} 映射到(0, 1)之间.

页面访问时间阈值的影响因子为

$$\beta=\left(R_{TCR}-R_{TCR_{\min}}\right)\Big/\left(R_{TCR_{\max}}-R_{TCR_{\min}}\right),$$

其中 $R_{TCR_{\min}}$ 为最大停留时间与 S_{DS} 页面大小的比值; $R_{TCR_{\min}}$ 为最小停留时间与 S_{DS} 页面大小的比值.

综合上述调整过程, 可以得出文字型阈值 δ 的计算公式为 $\delta=\alpha\, t(1+\beta)$.

1.4　基于页面停留时间阈值计算法

对媒体页面平均访问时间为 $R_{AVT}=\sum_{i=0}^{n}t_i/n.$

为了将 R_{AVT} 映射到(0，1)之间，考虑到孤立点的影响，同样采用定义2的计算公式

$$\beta = 1-\exp(-\mathrm{sqrt}(\mathrm{sqrt}(R_{AVT}))),$$

其中 β 为页面访问时间阈值的影响因子.

同理综合上述调整过程，可以得出阈值 δ 的计算公式为 $\delta=\alpha\, t(1+\beta)$.

1.5　识别方法的算法描述

要设置每个页面的访问时间阈值，首先要获得统计后的每个页面的访问时间t，然后根据不同URL类型计算影响因子 β 来调整 δ．统计后页面的访问时间t的集合记为 $S_t=(t_1,t_2,\cdots,t_n)$，页面的影响因子 β 集合记为 $S_\beta=(\beta_1,\beta_2,\cdots,\beta_n)$．用户会话集合记为 $S_{es}=(s_1,s_2,\cdots,s_n)$．算法步骤如下:

(1) 根据日志文件统计得到t的集合 S_t.

(2) 根据URL类型处理: (a)导航型URL采用链接内容比法; (b) 文字型URL采用基于页面内容与页面停留时间阈值计算法; (c) 多媒体型URL采用基于页面停留时间阈值计算法．根据以上条件采用不同的算法得到影响因子集合 S_β 和 α 值调整 S_t 得到页面访问时间阈值集合 S_δ.

(3) 根据 S_δ 重新划分日志文件得到用户会话集合 S_{es}.

该算法针对不同页面的特点，采用不同的方法来计算阈值影响因子，从而使结果更接近现实，会话的识别率高，但在运算效率上需要更多的资源，时间消耗在对页面的访问时间的计算和排序上，站点的页面数量越多，效率就越低.

2　实验与结果分析

2.1　数据预处理

选用 http: //www.pandawill.com 电子商务购物网站从 2009 年 11 月到 2010 年 1 月的 Web 服务器日志，共 12 万条记录, 101 871 个有效页面，其中对图片和不相关的数据进行剪裁，并根据链接地址格式标记 URL 类型.

得到数据表格式包含以下信息: 用户 IP，页面链接，来源链接，浏览器类型，页面大小，会话 ID，页面类型.

2.2　实验结果评估

实验如表 2 所示，发现 3 种不同 URL 类型的页面的时间阈值有较大不同，导航型 URL 的时间阈值普遍比文字要小，该现象与理论相符合．多媒体型的URL的时间阈值差异较大，通过分析，产生该现象的原因主要以图片为主的页面和与视频为主的页面差异较大，用户在一个视频页面的停留的时间要么很短，要么很长．产生该现象的原因是由视频页面的特点决定的.

表 **2**　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(R)	101 871	$\|R\|$=21 986	$\|R\bigcap R\|$=21 986	$\|R\bigcap R\|/\|R\|$=100	$\|R\bigcap R\|/\|R\|$=100
基于固定时间阈值(T)	101 871	$\|T\|$=40 362	$\|T\bigcap R\|$=10 360	$\|T\bigcap R\|/\|T\|$=25	$\|T\bigcap R\|/\|R\|$=47
基于页面访问时间阈值(A)	101 871	$\|A\|$ = 48 851	$\|A\bigcap R\|$=13 069	$\|A\bigcap R\|/\|A\|$=26	$\|A\bigcap R\|/\|R\|$=59
本文(L)	101 871	$\|L\|$ = 49 381	$\|L\bigcap R\|$=14 139	$\|L\bigcap R\|/\|L\|$=28	$\|L\bigcap R\|/\|R\|$=64

3　结束语

Web 日志挖掘具有广阔的发展空间和应用环境．通过精确地识别出用户和会话，才能发现用户的浏览模式．本文提出的会话识别改进算法，考虑不同页面类型访问差异等参数来计算用户对页面的浏览阈值，可以识别出页面浏览时间较长的会话，也可以把小于固定阈值的页面划分到下一个会话，从而可以更加个性化，更加准确地识别出会话．但它同样存在着不足，由于本文算法要求日志比较完整，如何处理数据域丢失的情况，是后续研究要考虑的问题；另外本文算法需要分析不同的页面类型，计算具体的页面阈值，如何提高改进的会话识别算法的效率也需要进一步优化.

4　参考文献

[1] 蔡浩，贾宇波，黄程伟. Web 日志挖掘中的会话识别算法 [J]. 计算机工程与设计, 2009, 30(6): 1321-1323.

[2] 方元康，胡学钢，夏启寿. Web 日志预处理中优化的会话识别

方法 [J]. 计算机工程, 2009, 35(7): 47-51.

[3] 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.

[4] 周爱武, 程博. Web 日志挖掘中的会话识别方法 [J]. 计算机工程与设计, 2010, 31(5): 936-938.

[5] 李燕, 冯博琴. Web 日志挖掘中的数据预处理技术 [J]. 计算机工程, 2009, 35(22): 44-49.

[6] 范纯龙, 姜宏飞. 利用图片类日志信息改进会话识别质量 [J]. 计算机应用, 2010, 30(4): 1056-1058.

[7] 杨富华. 网络日志预处理中优化的会话识别算法 [J]. 计算机仿真, 2011, 28(4): 123-125.

[8] 方元康, 王汝传. 优化的 Web 日志会话识别方法 [J]. 计算机工程与设计, 2009, 30(7): 1688-1690.

[9] Spiliopoulou M, Mobasher B, Berendt B, et al. A framework for the evaluation of session reconstruction heuristics in web usage analysis [J]. Informs Journal of Computing, 2003, 15(2): 171-179.

[10] Facca F M, Lanzi P L. Mining interesting knowledge from web logs: a survey [J]. Data and Knowledge Engineering, 2005, 53(3): 225-241.

[11] Sucitanek F M, Ifrim G, Gerhard W, et al. Combining linguistic and statistical analysis to extract relations from Web documents [C]. New York: ACM, 2006: 712-717.

[12] He Xinhua, Wang Qiong. Dynamic timeout-based a session identification algorithm [EB/OL].[2012-01-12].http://ieeexplore.ieee.org/xpl/article Details.jsp?arnumber=5777587.

Improved Method of Session Identification Based on Page and Time Threshold

ZHENG Li-shan, TENG Shao-hua*

(Faculty of Computer, Guangdong University of Technology, Guangzhou Guangdong 510006, China)

Abstract: Based on the commonly used method of computing time threshold method, an improved method of session identification which based on page type, page size, visiting time is brought forward. For different page types, different threshold calculation methods are used to set the time threshold. Relative to the existing use of a single priori threshold and current dynamic threshold computing method, the method can give more realistic reflection of the session situation and the accuracy has been greatly improved.

Key words: Web mining; data preprocessing; user session identification; dynamic threshold

(责任编辑: 冉小晓)