基于聚类的 Web 用户会话识别优化方法*

凌海峰[1,2]，余 笪[1,2]
(1. 合肥工业大学 管理学院，合肥 230009；2. 过程优化与智能决策教育部重点实验室，合肥 230009)

摘 要：会话识别是用户访问行为分析的基础和关键工作，其质量对于识别和发现用户的信息需求具有决定性的影响。目前常用的是基于时间阈值的切分方法，但是该方法存在的主要问题是针对不同用户时间阈值难以准确地确定。提出了一种新的基于聚类技术的会话识别优化方法，首先建立了基于聚类的会话识别优化模型，然后采用改进的 K-means 算法进行会话识别。实验结果表明该方法与传统方法相比具有较好的效果。

关键词：Web 用户会话识别；K-means 算法；时间阈值
中图分类号：TP391　　**文献标志码**：A　　**文章编号**：1001-3695(2012)08-2862-03
doi:10.3969/j.issn.1001-3695.2012.08.016

Optimized methodology for Web user session reconstruction based on clustering

LING Hai-feng[1,2], YU Da[1,2]
(1. *School of Management*, *Hefei University of Technology*, *Hefei* 230009, *China*; 2. *Key Laboratory of Process Optimization & Intelligent Decision-making of Ministry of Education*, *Hefei* 230009, *China*)

Abstract: As the basic and critical work for user access behavior analysis, Web user session reconstruction has a decisive impact on identifying and discovering the information needs of users. Currently Web user session reconstruction is usually based on the time threshold segmentation method. But it is difficult for this method to accurately determine the time threshold for different user. This paper presented a new method for Web user session reconstruction based on clustering method. Firstly, it created an optimization model of Web user session reconstruction based on clustering, and then used an improved K-means algorithm for this model. Experimental results show that this method has better results compared with traditional methods.
Key words: Web user session reconstruction; K-means algorithm; time threshold

0 引言

随着互联网特别是电子商务的快速发展，消费者的网络访问行为对于电子商务中购买预测模型的建立、用户访问行为的分类等显示出极端的重要性。而 Web 访问日志体现了消费者使用 Web 资源的行为特点和规律，是消费者访问行为分析最重要的信息源之一[1,2]。会话识别是用户访问行为分析的基础和关键工作，会话识别的准确度对于识别和发现用户的信息需求具有决定性的影响。会话是指用户在一次访问过程中所访问的页面序列，它代表了用户对服务器的一次有效访问。会话识别的目的就是将每个用户访问的页面划分到不同的会话中，以会话为基本单元将有助于模式的挖掘和分析。

目前会话识别大多数还是基于启发式算法，大体可分为以下两类[3]：

a)基于时间阈值的启发式算法。Catledge 等人[4]通过实验度量了用户对一个站点的一次访问持续时间，将 30 min 作为一次会话的最长持续时间；He 等人[5]通过实验给出了时间阈值，认为将页面停留的时间阈值确定在 10～15 min 是最优的会话时间间隔；Yan 等人[6]统计每个用户的频率矢量，在正态分布的假设下，为每一个用户设定一个合理的切分阈值；殷贤亮等人[7]根据页面内容及站点结构重要程度来确定阈值；Huynh 等人[8]则提出基于实验观察超时阈值模型。

b)基于引用的启发式算法。周爱武等人[9]以 Web 日志中用户访问页面序列的 URL 作为划分会话的标准。由于不同的用户和不同的网页具有各自的特点，此类启发式方法明显地具有主观化和机械化等缺点，从而使得不同用户的时间阈值难以准确地确定。还有一些学者开始尝试采用其他方法来处理会话识别问题。张辉等人[10]采用 Markov 链模型识别会话；Dell 等人[11]考虑网页中回退按钮的情况，从运筹学的角度利用 0-1 整数规划进行会话识别，该方法的优点在于可以一次性构建所有的会话，但在面对大量的网络数据时需要花费巨大的时间成本。

本文从聚类优化的视角研究 Web 会话识别方法，主要研究基于页面时间距离的会话识别自适应优化方法。基本思想就是：如果两个连续页面的访问时间间隔越大，则它们属于相同会话的可能性就越小。K-means 算法[12]具有可伸缩性和效率高的优点，本文提出改进的 K-means 算法来克服传统算法只可应用于无序的聚类情况，从而解决目前会话识别过程中所面临的问题。实验结果表明，与传统的时间阈值方法相比，该方法具有较好的会话识别效果。

1 基于聚类的会话识别优化模型

假设在预处理过程中的用户识别之后，得到某个用户的日

收稿日期：2012-01-13；**修回日期**：2012-02-21　　**基金项目**：国家自然科学基金资助项目(71071047)；高等学校博士学科点专项科研基金资助项目(20090111110016)；安徽省自然科学基金资助项目(1208085MG120)；合肥工业大学博士专项基金资助项目(2010HGBZ0301)

作者简介：凌海峰(1971-)，男，安徽肥东人，副教授，博士，主要研究方向为电子商务、智能优化(hfling@126.com)；余笪(1986-)，男，安徽桐城人，硕士，主要研究方向为电子商务、数据挖掘.

志记录中包含了 m 个 URL, 即 $U=\{\text{url}_1,\text{url}_2,\cdots,\text{url}_m\}$,它包含了 k 个会话。首先给出该问题的形式化定义。

定义 1　一个用户会话可以表示为 session = 〈sessionID, $\{(\text{url}_1,\text{time}_1),(\text{url}_2,\text{time}_2),\cdots,(\text{url}_n,\text{time}_n)\}$〉,$n\leqslant m$。其中:sessionID 为会话标志;url 为用户访问的网页;time 为服务器响应网页请求的时间。

定义 2　会话识别就是将一个用户对 Web 服务器的一次有效访问进行识别。在这个会话内所访问的网页在内容上是相关的。假设某个用户的 Web 日志中包含 m 个 URL,即 $U=\{\text{url}_1,\text{url}_2,\cdots,\text{url}_m\}$,它包含了 k 个用户会话,即 $S=\{S_1,S_2,\cdots,S_k\}$; $\text{session}S_j$ = 〈sessionID, $\{(\text{url}_1,\text{time}_1),(\text{url}_2,\text{time}_2),\cdots,(\text{url}_n,\text{time}_n)\}$〉,其中 $S_j\in S$。

本文将采用优化的方法来划分用户会话。解决此类问题的目标是将某一个用户的 m 个日志记录最优地分配到 k 个聚类中,使得每个日志记录与所属类的质心之间的欧氏距离最小。创建的解的质量是由该问题的目标函数值来度量。

现给定某个用户的日志记录包含 m 个日志记录的数据集,将其划分为 k 组,则该划分问题的数学模型可以描述为

$$\min F(w,a_i,b_j)=\sum_{j=1}^{k}\sum_{i=1}^{m}w_{i,j}\parallel a_i-b_j\parallel^2$$

并满足

$$\sum_{j=1}^{k}w_{i,j}=1\quad i=1,\cdots,m \tag{1}$$

$$\sum_{i=1}^{m}w_{i,j}\geqslant 1\quad j=1,\cdots,k \tag{2}$$

其中:$w_{i,j}$是指日志记录 i 相对簇类 j 的权重,且

$$W_{i,j}=\begin{cases}1 & \text{如果日志记录 } i \text{ 属于聚类 } j\\ 0 & \text{否则}\end{cases}$$

式(1)确保每条记录只存在于某一个会话之中;式(2)确保每一簇类至少包含一条日志记录。

2　基于改进 K-means 的会话识别优化算法

会话识别在 Web 使用挖掘中占有极其重要的一步。很多学者都在研究如何提高会话识别的精确度和效率。但是由于一个网站的 Web 服务器所记录的数据量非常庞大,使得很多方法的精确度或效率都不是很理想。K-means 算法具有可伸缩性和效率极高的优点,因而被广泛地应用于大文档集的处理。而传统的 K-means 算法只是针对无序的数据进行聚类,并没有考虑数据间的有序性。

本文提出一种改进的 K-means 算法来处理具有有序数列特征的时序数据的聚类问题。与传统方法相比,优化的方法不仅在准确度上有了一定的提高,而且通过优化方法可以一次将所有的会话全部识别出来。其主要思想就是在聚类的过程中按照数据间的时间顺序来逐步聚类。具体流程步骤如下:

a)将用户识别后的 Web 日志记录按照 Web 服务器响应时间顺序进行编号。

b)确定时间阈值。

c)确定初始聚类中心点和 k 值。

d)将得到的初始聚类中心点按照时间顺序进行编号。

e)按照日志记录和聚类中心点的编号顺序进行聚类。

f)计算新簇类的均值,更新聚类中心点。

g)重复迭代步骤 e)f),直到聚类中心点不再发生改变为止。

以下本文就分别详细介绍流程步骤中关键的 b)c)和 e)这三步。

2.1　时间阈值的确定

本文采用一定的时间阈值的方法来确定 K-means 算法的 k 值和初始聚类中心点的位置。在确定每个用户的时间阈值时,采用每个用户的平均页面浏览时间 μ 和方差 3σ 为该用户的初始聚类的时间阈值。在数据的处理过程中,首先在用户识别的基础上计算每一个用户的平均页面停留时间和时间的方差,由此得出该用户的时间阈值 $\theta=\mu+3\sigma$。当页面停留时间不超过一个提前计算得到的阈值 θ 时,则将该记录划分在同一簇中,否则划分到不同簇中。

由于服务器只能记录每个请求的响应时间,所以用下一个页面响应时间减去当前响应时间的结果为该请求页面的浏览时间。在计算平均页面停留时间时,在考虑一个用户的最后一个网页的页面停留时间时,使用该用户的平均时间作为该页面的浏览时间。

通过这种方法得出每一个用户自己特有的阈值,从而在一定程度上减少了以前 30 min 切分会话的不足。

2.2　确定 k 值和初始聚类中心

在 K-means 算法中,k 值表示聚类的簇数,其值在很大程度上决定算法聚类的精确度;如何选择初始聚类中心点也是影响最后聚类结果的重要因素。因此,如何确定 k 值和初始聚类中心成为 K-means 算法亟待解决的重要问题。传统的K-means 算法的 k 值是根据经验事先确定好或者随机产生的,这样做有很多不足,其中最主要的缺陷是由于不同的用户的个人兴趣和每个网页都具有自身的特点。该方法在确定 k 值的过程中添加了很多的个人主观意愿因素,具有很大的主观性和机械性,不能根据数据的客观情况来准确地聚类。

本文根据 Web 日志记录中的客观现实情况来计算并动态调整 k 值和初始聚类中心位置。首先根据数列的顺序从小开始计算与它相邻数列的距离。若两者的距离小于一个事先确定的阈值 $\theta=\mu+3\sigma$,则该两个数列可以划分到同一个类别中;否则将它们划分到不同的会话中。依此类推,不断循环,从而计算出初始簇类的个数和初始聚类中心点的位置,k 值即为初始簇类的个数。

具体过程如下:

a)将 Web 日志序列每个用户记录按照时间顺序进行编号,如 $a_1,a_2,\cdots,a_m$(m 为序列总个数)。

b)计算相邻的两个日志记录的距离 $d(a_i,a_{i+1})$。若 $d(a_i,a_{i+1})\leqslant\theta$,则将第 i 条记录和第 $i+1$ 条记录归为同一簇类;否则将它们划分到不同的簇类中。

c)重复循环步骤 b),直到将所有的 Web 日志记录完全划分到不同的簇类之中。

d)计算得出初始聚类的簇类个数,即为 k 值。对于最终划分的簇类,计算每一簇类的均值,以此为该簇类的初始聚类中心点。

2.3　聚类过程

具体的 K-means 聚类步骤如下:

a)依次将时间序列上的有序数列进行编号,如 $a_1,a_2,\cdots,a_m$(m 为序列总个数)。按照 2.2 节的步骤选取 k 个观测值作为初始聚类中心,并将其按照时间顺序依次编号,如 $b_1,b_2,\cdots,$

b_k。将距离 a_1 最近的初始聚类中心点记为 b_1，且 $b_i \in S_i$（$i \leqslant k$）。

b）按照聚类中心的顺序排列每次比较相邻的某一个数列到聚类中心的距离 $d(a_i, b_j)$，表示第 i 个序列到第 j 个聚类中心的欧氏距离，其中 $i \leqslant m, j \leqslant k$。若 $d(a_i, b_j) \geqslant d(a_i, b_{j+1})$，则 $a_i \in S_{j+1}$，反之 $a_i \in S_j$。

c）对于重新划分好的各簇类，选取每一簇类的平均时间为该簇类新的聚类中心点来替代上一步的聚类中心点。

d）重复步骤 b）c），直至达到与上一步的聚类中心点没有发生改变为止。

在此迭代步骤过程中，与传统的 K-means 算法中要计算每个节点到所有的聚类中心的距离相比，改进的 K-means 算法极大地降低了算法的时间复杂度。由于数列的时间顺序性，使得在聚类过程中都是按照此顺序进行迭代。在一次迭代结束后得出 $a_i \in S_j$。令 b_j 是 S_j 的聚类中心点，则在下一迭代过程中，只需要比较 $d(a_i, b_{j-i})$，$d(a_i, b_j)$ 和 $d(a_i, b_{j+1})$ 三者之间的关系即可，无须再计算比较 a_i 与其他的聚类中心点的距离。

3 实验结果

3.1 评价标准

本实验中，利用类内致密性和类间离散性来构建会话识别的评价标准。

a）类内致密性。设一个用户记录中包含 m 个记录，并且可以将其划分为 c 个聚类，每个聚类中含有 n_i 个样本（n_i 表示第 i 个聚类中包含的记录个数），则聚类结果的类内致密性度量定义为

$$R_{\text{in}} = \sum_{i=1}^{c} \left(\frac{1}{n_i} \sum_{j=1}^{n_i} \| a_j - a^{(i)} \| \right)$$

b）类间离散性。设一个用户记录中包含 m 个记录，并且可以将其划分为 c 个聚类，每个聚类中含有 n_i 个样本，则聚类结果的类间离散性度量定义为

$$R_{\text{be}} = \frac{1}{m} \sum_{i=1}^{c} n_i \| a^{(i)} - \bar{a} \|$$

其中：$a^{(i)}$ 表示第 i 簇类的聚类中心点，$\bar{a}$ 表示所有簇类的中心点。

根据类内致密性度量和类间离散性度量来构建新的有效性指标 V：

$$V = R_{\text{in}} / R_{\text{be}}$$

V 值越小说明结果越好。

3.2 实验结果

本实验的分析平台是酷睿 2 双核处理器，2 GB 内存，500 GB 硬盘，宏基 Veriton M275 台式电脑，操作系统为 Windows XP，在 MATLAB 7.0 的环境下编写程序。实验数据来自 EPA-HTTP 日志，它包含用户在 24 h 内访问 EPA 服务器的请求，总大小为 4.24 MB。经数据清洗和用户识别后得到8 253条有用的日志记录，删除只有一次访问记录的用户之后得到1 121个用户。

本实验分别进行了三组实验，将改进后的方法与传统的基于 30 min 的会话划分和基于 $\theta = \mu + 3\sigma$ 的会话划分[13] 相比较。表 1 显示了会话识别的会话个数以及评价指标结果。从表 1 中可以看出，与传统方法相比，本文提出的基于聚类的会话识别方法具有更高的准确度。

表 1 会话识别结果的比较

比较项	会话个数	R_{in}	R_{be}	$R_{\text{in}}/R_{\text{be}}$
30 min	1 532	4 371.8	892.850 5	4.896 5
$\theta = \mu + 3\sigma$	2 539	4 386.1	931.133 2	4.710 5
改进算法	2 539	4 390.6	956.748 2	4.589 1

4 结束语

本文给出了一种基于改进 K-means 算法的会话识别优化方法。实验结果表明该方法是有效的，与传统的基于时间阈值方法相比，该方法显著地改善了会话识别的准确度。

笔者今后的研究方向主要是如何进一步提高会话识别的准确度，可以从以下两个方面进行：

a）在聚类过程中考虑页面间距离时只考虑页面间的时间差这一维特征，在会话识别过程中具有一定的局限性。在以后的研究过程中，可以将用户浏览的网页 URL 的内容作为另一维特征来提高会话识别的准确度。

b）在使用 K-means 算法过程中，k 值确定方法仍然参照了时间阈值，如何自动确定有效的 k 值，在很大程度上会影响到算法的准确度。

参考文献：

[1] 孙宇航. 基于网络日志的数据挖掘预处理改进方法[J]. 系统工程与电子技术，2009，31(12)：2994-2997.

[2] 方元康，胡学刚，夏启寿，等. 改进的 Web 日志数据预处理技术[J]. 计算机工程，2009，35(10)：73-77.

[3] PRASAD G S，REDDY N V S，ACHARYA U D. Konwledge discovery form Web usage data：a survey of Web usage pre-processing techniques[C]//Proc of International Conference on Recent Trends in Bussiness Administration and Information Processing. 2010：505-507.

[4] CATLEDGE L，PITKOW J. Characterizing browsing behaviors on the world wide Web[J]. Computer Networks and ISDN Systems，1995，26(6)：1065-1073.

[5] HE Da-qing，GOKER A. Detecting session boundaries from Web user logs[C]//Proc of the 22nd Annual on Information Retrieval Research. 2000：57-66.

[6] YAN T W，JACOBSEN M，GARCIA-MOLINA H，*et al*. From user access patterns to dynamic hypertext linking[C]//Proc of the 5th International WWW Conference on Computer Networks and ISDN System. 1996：1007-1014.

[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，2006，34(7)：33-35.

[8] HUYNH T，MILLER J. Empirical observations on the session timeout threshod[J]. Information Processing and Management，2009，45(5)：512-528.

[9] 周爱武，程博，李孙长，等. 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)：936-938，964.

[10] 张辉，宋瀚涛，徐晓梅. 基于语义的 Web 用户会话识别算法[J]. 北京理工大学学报，2007，27(6)：471-473.

[11] DELL R F，ROMAN P E，VELASQUEZ J D. Web user session reconstruction with back button browsing[C]//Proc of the 13th International Conference on Knowledge-based and Intelligent Information and Engineering Systems，Part Ⅰ. Berlin：Springer-Verlag，2009：326-332.

[12] 周世兵，徐振源. 新的 K-均值算法最佳聚类数确定方法[J]. 计算机工程与应用，2010，46(16)：27-31.

[13] 庄力可，寇忠宝，张长水. 网络日志挖掘中基于时间间隔的会话切分[J]. 清华大学学报：自然科学版，2005，45(1)：115-118.