第 26 卷 第 3 期
2011 年 6 月

天津科技大学学报
Journal of Tianjin University of Science & Technology

Vol.26 No. 3
Jun. 2011

基于点击流的频繁模式聚类算法研究

李 杨，檀柏红
（天津科技大学经济与管理学院，天津 300222）

摘 要：在用户访问网站点击流形成频繁序列的基础上，提出基于距离函数的聚类分析算法. 首先对数据流分区做 K 均值聚类生成中间聚类结果，然后对这些均值参考点进行离线聚类，以获取用户访问模式. 理论分析和实验表明，算法具有较好的聚类效果.

关键词：点击流；聚类；频繁模式
中图分类号：TP391.4 **文献标志码**：A **文章编号**：1672-6510(2011)03-0069-05

Clustering Algorithm of Web Click Stream Frequency Pattern

LI Yang, TAN Bai-hong
(College of Economics and Management, Tianjin University of Science & Technology, Tianjin 300222, China)

Abstract: Base on the frequency sequence pattern by the Web click flow frequency constitutes, the analysis of the clustering algorithm according to the distance function was put forward. K means clustering was adopted on each partition of the data stream to generate mean reference point set, and subsequently offline clustering was applied to get the clustering result of each periods to obtain the user access pattern. Theoretic analysis and experiments show that the algorithm is effective and efficient.

Keywords: click stream; clustering; frequency pattern

随着网络流量的激增和电子商务的快速发展，怎样通过用户与网站的交互了解其喜好与浏览习惯是网站运营者迫切需要解决的问题.

聚类分析是数据挖掘领域的一项重要研究课题. 聚类算法的目的是将数据对象自动地归入到相应有意义的簇中，追求较高的簇内相似度和较低的簇间相似度.

点击流数据是一种典型的流数据，除了具有流数据的特性，点击流还具有自身特点. 对点击流聚类的研究有诸多关键问题没有被最终解决，其中 2 个问题是：(1)点击流的原始数据是以序列的形式而不是特征向量的形式存在，许多对特征向量聚类效果良好的聚类算法不能直接使用[1]；(2)点击流数据量巨大，致使许多聚类算法效果不好[2]. 针对上述问题，本文提出一种点击流聚类方法 CluClick，算法通过对点击流模式间的相似性进行距离度量的定义，并借鉴经典算法 CluStream 分为在线层和离线层两个框架.

1 点击流数据描述

用户连接到一个网站时，服务器会在日志文件中记录其所有操作，其一系列的访问过程形成 Web 点击流，Web 点击流数据记录了访问者访问 Web 站点的过程. 使用者每次按某一顺序访问一系列 Web 页面的过程被称作是一个会话(session)，Web 网站的点击流数据是发生于该网站的一组会话[3]，记录 Web 点击流数据的文件或数据库数据被称作用户访问日志. 日志文件中的每个记录包含客户端的 IP 地址、收到请求时间、请求对象等信息.

从日志文件中可以提取用户访问模式. 但通常需要做一些诸如数据清理和会话识别的数据预处理工作[4]. 文献[5]描述了如何在 IIS Web 服务器中提取

收稿日期：2011–01–07；**修回日期**：2011–03–08
基金项目：天津市科技发展战略研究计划项目(10ZLZLZF4900)
作者简介：李 杨(1980—)，男，山东人，讲师，erlee@tust.edu.cn.

Web 日志进行数据预处理，并介绍了相关处理技术. 为便于分析，把一个服务器会话作以下处理：用字母表代替一组网页，每个字母对应一个网页，用户浏览网页的序列可用该字母表的字符串表示. 于是作如下处理：令字母表 $B=\{b_1,b_2,\cdots,b_n\}$ 为某一站点的一组网页，$s=a_0,\cdots,a_i,\cdots,a_k(a_i\in B,0\leqslant i\leqslant k,k\geqslant 0)$ 是一用户在登录此站点后浏览的网页序列，称 s 为一个点击流或一个服务器会话.

2　类紧密程度的度量

一个用户会话定义为时间紧凑的用户访问序列. 由于先验知识缺乏，聚类方法非常适合进行用户会话分析. 目前，主要有两种应用于点击流的聚类，一种是把点击流看成是对象数据，另外一种则以点击流的相似性为基础. 本文采用了基于相似性的聚类方法.

由于点击流的流数据特性，仅定义访问事务间的相似性，将访问事务看作是数据点来进行一次聚类，不能满足将动态聚类结果和原有聚类合并的要求. 因此需要进一步定义类内紧密程度和类间距离.

文献[6]中给出了一种基于频繁模式的距离定义，假设 P_1 和 P_2 为频繁模式，$T(P_1)$ 和 $T(P_2)$ 分别为 P_1 和 P_2 的支持度列表，则 P_1 和 P_2 之间的距离为

$$D(P_1,P_2)=1-\frac{|T(P_1)\bigcap T(P_2)|}{|T(P_1)\bigcup T(P_2)|}$$

距离 D 有如下性质：

(1) $D(P_1,P_2)>0,\forall P_1\neq P_2$

(2) $D(P_1,P_2)=0,\forall P_1=P_2$

(3) $D(P_1,P_2)=D(P_2,P_1)$

(4) $D(P_1,P_2)+D(P_2,P_3)\geqslant D(P_1,P_3),\forall P_1,P_2,P_3$

此度量函数实际反映了两个访问模式之间的相似程度，D 越小则表明 P_1 和 P_2 两模式越相似，反之则差异性越大. 该距离度量的意义在于，基于这样的距离度量充分考虑了用户访问模式，并使得多数距离算法可以应用.

2.1　类内紧密程度的度量

给定一个聚类 $C=\{c_1,c_2,\cdots,c_p,\cdots,c_m\}$，并且 c_p 是聚类的代表事务，则聚类 C 的类内紧密程度定义为

$$dist\overline{C}=\frac{1}{m}\sum_{c_i\in C}D(c_i,c_p)$$

式中：$D(c_i,c_p)$ 是 c_i 到参考点 c_p 的序列模式距离. 此紧密程度数值越小则表明类内所包含模式越趋于一致.

对于任意一个会话事务，包含了用户访问的路径及其驻留时间. 因此，对访问事务进行聚类需要定义其距离度量，其距离度量涉及到访问页面的兴趣度.

2.2　访问页面兴趣度

I_l 定义为访问事务中对页面 l 的兴趣：

$$\begin{cases}I_l=0,t<t_{\min}\\ I_l=(n-l+1)\times\dfrac{t-t_{\min}}{t_{\max}-t_{\min}}\times f(l),t_{\min}\leqslant t\leqslant t_{\max}\\ I_l=(n-l+1)\times f(l),t>t_{\max}\end{cases}$$

式中：n 为事务长度；l 表示本页面在事务中的位置；$f(l)=\dfrac{s_l}{s_{\text{all}}}$，$s_l$ 为第 l 个页面在检测时间内访问次数，s_{all} 为该事务中所有页面的访问总次数；t 为访问时间，$t_{\min}$ 和 $t_{\max}$ 分别为访问时间控制阈值，小于 $t_{\min}$ 为无效访问，大于 $t_{\max}$ 不增加兴趣度. 该兴趣度度量以访问时间为用户访问页面的兴趣标准，并且结合访问结构强调页面的访问频繁度.

2.3　访问事务的距离度量

给定 s^k 和 s^l 两个访问事务，其距离度量如下：

$$d(s^k,s^l)=(1-\frac{\sum_{i=1}^{N}I_i^kI_j^l}{\sqrt{\sum_{i=1}^{N}I_i^k}\sqrt{\sum_{i=1}^{N}I_i^l}})^2$$

式中：N 为页面总数；$I^k{}_i$ 和 $I^l{}_i$ 为事务 s^k 和 s^l 对页面 i 的兴趣度.

2.4　类间紧密程度的度量

给定两个聚类 C_1 和 C_2，类间紧密程度采用平均距离表示，若有 $P_i\in C_1,Q_j\in C_2$，则 C_1 与 C_2 的类间距离定义为

$$dis(C_1,C_2)=\frac{1}{n_1n_2}\sum_{P_i\in C_1,Q_j\in C_2}d(P_i,Q_j)$$

式中：n_1,n_2 分别为聚类 C_1 和聚类 C_2 所包含访问事务的数量；$d(P_i,Q_j)$ 为访问事务 P_i 和 Q_j 的距离度量.

类间距离即两聚类各自包含的数据点间的平均距离，用于考察两个聚类之间的相似程度. 对于任意两聚类，该值越小则表明两聚类越趋向于一致，在后续处理中将其并入统一聚类的可能性越大.

对于任意两个聚类，根据定义确定类内距离和类间距离，如果类间距离小于两聚类内聚类，则两聚类

应合并为同一聚类. 该过程不断重复直至找不到符合该条件的任意两聚类.

3　基于流数据的算法改进

3.1　改进原理

对于点击流聚类来说，流数据可以进行如下描述[7]：设 E 为网站所包含的所有网页的集合，则一次会话的点击流序列可描述为 $S=<e_1, e_2, \cdots, e_N>$，其中，$e_i \in E$，$i=1, \cdots, N$，是用户的一次点击事件，$N$ 为该次会话所持续的序列长度.

要进一步挖掘的对象就是由所有点击流序列组成的点击流序列集合 $SeqSet=\{S_1, S_2, \cdots, S_n\}$，其中 n 为不同的会话事务即点击流序列的个数. 算法的改进基于上节中的距离度量定义，在距离度量的基础上通过 K 均值算法实现对访问事务的聚类，并且通过对类内和类间距离的定义基于访问频繁模式对聚类进行调整.

对于点击流，数据流中数据块以点击流序列集合 $SeqSet_1, SeqSet_2, \cdots, SeqSet_m$ 形式到达，每个数据块包含 n 个序列. 基于定义 2 和定义 3 的距离度量，算法对每一个数据块使用 K 均值算法进行聚类，数据块划分成 p 个聚类，记为 $SeqSet=C_1 \cup C_2 \cup \cdots \cup C_p$，并生成 p 个代表模式（即参考点），记为 $c_1, \cdots, c_i, \cdots, c_p$. 对于每一个聚类 C_i，$i=1, \cdots, p$，记为 $C_i(c_i, s_i, d, n_i)$，其中 $s_i=dist\overline{C}_i$，x_j 为隶属于参考点 c_i 的序列，d 为聚类中距离 c_i 的最远距离，n_i 是隶属于 c_i 的序列数目.

3.2　聚类合并度量

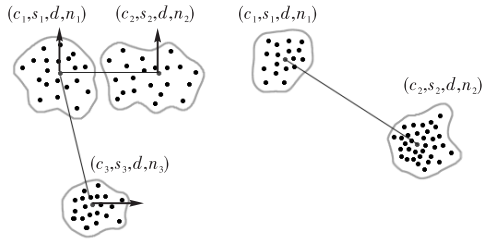
基于上述定义的距离度量，对于任意两参考点所代表的聚类 $C_1(c_1, s_1, d, n_1)$ 和 $C_2(c_2, s_2, d, n_2)$，若满足如下条件：

(1) $1-\beta \leqslant dist\overline{C}_1 / dist\overline{C}_2 \leqslant 1+\beta$

(2) $dis(c_1, c_2) \leqslant d_1+d_2$

其中，β 为阈值参数，则 $C_1(c_1, s_1, d, n_1)$ 和 $C_2(c_2, s_2, d, n_2)$ 可被合并.

上述定义参考文献[8]中的密度相邻描述形式，但其统一使用欧几里德距离进行的密度衡量不适用于点击流的事务序列聚类. 有别于该文献，此处定义条件(1)以序列模式相似性进行衡量，条件(2)采用基于点击流的页面兴趣度和访问事务距离度量进行衡量. 如图 1(a)所示，$C_1(c_1, s_1, d, n_1)$ 和 $C_2(c_2, s_2, d, n_2)$ 相对于 $C_3(c_3, s_3, d, n_3)$，其类间相似度高且类内相似度差异不大，因此在聚类过程中倾向于合并. 而在图 1(b)中，$C_1(c_1, s_1, d, n_1)$ 和 $C_2(c_2, s_2, d, n_2)$ 类间距离大而类内相似性高则认为是满意的聚类.



(a)不满意的聚类　　(b)满意的聚类

图 1　距离关系图

Fig.1　Example of distance relations

4　算法描述

算法借鉴经典算法 CluStream 分为在线层和离线层两个框架. CluStream[9]是 Aggarwal 在 2003 年提出的一个解决数据流聚类问题的框架. 在线层快速接收输入的数据流，其产生的结果作为中间结果被维护起来，并且随着新数据点的流入该中间结果实时动态更新. 依据一定的时间框架周期性地选取某些特定时刻的中间结果，保存到外存作为离线层算法的输入. 离线层由用户调用，针对用户的挖掘请求，利用金字塔时间框架给出其感兴趣的不同时间粒度上的结果[10].

4.1　在线过程算法

输入：序列模式集 $P=\{a_1, a_2, \cdots, a_m\}$

输出：k 个序列模式聚类 $C_1, C_2, \cdots, C_k$

(1)读入经过积累一段时间的模式集；

(2)根据定义距离度量计算；

(3)对数据基于 K 均值算法进行聚类，聚类结果形成互不相交的聚类块；

(4)读入新的模式；

(5)根据聚类度量确定该模式到各聚类块中心点的距离，以此为标准搜索是否存在一个包含该模式的聚类块；

(6)如果存在包含该模式的聚类块则指派该模式给离他最近的中心点所代表的簇，否则创建为新聚类块；

(7)输出 k 个序列模式聚类块为中间结果.

4.2　离线过程算法

输入：时间 t_1, t_2，阈值 E，k 个聚类 $C_1, C_2, \cdots, C_k$

输出：聚类结果

(1) 对 t_1 到 t_2 范围内的聚类块进行检查;

(2) 如果在阈值范围内有相似类,将其聚合聚类块,并更新聚类特征;

(3) 如果一个聚类块找不到相似簇,则将其作为新簇;

(4) 重复上述过程直到没有新的聚类块可以处理;

(5) 输出聚类结果.

5　实　验

为评估算法的质量和效率采用聚类纯度[11]作为聚类评价指标. 聚类纯度定义为[12],每个簇中占主导地位的类别元素数目与簇的大小比值的加权平均值.

实验在配置为 Intel Core(TM)2 duo CPU E8300 2.83 GHz, 4.0 GB RAM 的 PC 上进行,操作系统为 Windows XP 专业版,算法用 VC++6.0 实现. 测试数据集为从一个小型的电子商务网站获取的一个真实数据集. 该数据集包括 8 000 个点击事件,数据集描述了网站访问用户在该网站的 43 个网页上的浏览路径. 由于真实数据集中包含部分长度过短的无意义事务,经过数据预处理筛选出 1 492 条有效会话事务,每条事务对应用户对网站的一次浏览路径,平均每条事务包含 6.3 个点击事件.

当聚类纯度较高时,表明聚类结果较好,可以认为数据被恰当地归入了一个类别. 为了测试算法,在完成数据预处理与距离度量定义的基础上与经典的流数据处理算法 CluStream 进行比较. 为便于分析,将原数据每 100 条记录划分为一个数据块,间隔固定时间输入一个数据块,分别在不同时刻根据聚类结果计算聚类纯度. 在进行聚类的过程中,既考查聚类的类内相似度,同时也考查聚类的类间相似度,图 2 为聚类效果比较. 可以看出,在多数时间本文算法的聚类效果好于 CluStream(k = 6, 阈值 = 0.2). 因为算法利用 K 均值方法对聚类生成聚类块,并将参考点记录为四元组的形式,使得数据聚类记录了较为充分的数据分布信息.

聚类结果的分布见图 3,横坐标为各聚类结果,纵坐标为所属聚类的样本数. 从图 3 可以看出,近一半的样本都集中在第一类中,分布悬殊较大. 这是由于该类访问通常集中在与网站主页紧密相关页面,如介绍页面等. 其他事务所涉及页面类别也基本与页面内容相关.

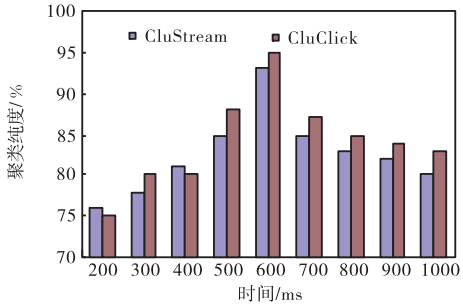


图 2　聚类效果比较

Fig.2　Effect comparison between CluStream and CluClick

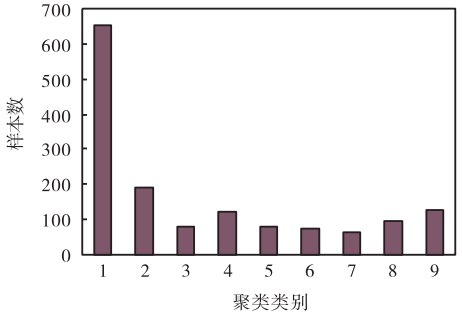


图 3　聚类结果分布

Fig.3　Distribution of cluster result

6　结　语

本文通过对用户访问事务的距离定义,解决了点击流数据点序列形式聚类的问题,给出了一种基于 Web 点击流的用户访问模式聚类算法. 为满足流数据的大数据量和实时性需求,算法通过线上和线下计算相结合的方式提高了算法的效率. 对于点击流数据,算法可用于研究点击序列的相关性和会话事务所反应出来的兴趣相似性,可以发现用户使用模式,从而对用户归类. 这对与用户个性化服务和网站页面合理布局都具有重要意义. 理论分析和实验结果表明,算法具有较好的聚类效果.

参考文献:

[1] Banerjee A, Ghosh J. Clickstream clustering using weighted longest common subsequences[C]// Proceedings of the Web Mining Workshop at the First SIAM International Conference on Data Mining. Chicago, USA: National Center for Data Mining, 2001: 33–40.

[2] Park S, Suresh N C, Jeong B K. Sequence-based clustering for Web usage mining: a new experiment al framework and ANN-enhanced K-means algorithm[J]. Data & Knowledge Engineering, 2008, 65(3): 512–543.

[3] Lee J, Podlaseck M, Schonberg E, et al. Visualization and analysis of click stream data of online stores for understanding Web merchandising[J]. Data Mining and Knowledge Discovery, 2001, 5(1/2): 59–84.
[4] Cooley R, Mobasher B, Srivastava J. Data preparation for mining World Wide Web browsing patterns[J]. Knowledge and Information Systems, 1999, 1(1): 5–32.
[5] Mohd Helmy Abd Wahab, Mohd Norzali Haji Mohd, Hafizul Fahri Hanafi, et al. Data pre-processing on Web server logs for generalized association rules mining algorithm[J]. World Academy of Science, Engineering and Technology, 2008, 48: 190–197.
[6] Xin D, Han J, Yan X, et al. Mining compressed frequent-pattern Sets[C]// Proceedings of the 31th VLDB Conference. Trondheim, Norway: VLDB Endowment Inc., 2005: 709–720.
[7] 马超，沈微. 基于闭合有间隔频繁子序列的点击流聚类[J]. 计算机工程，2010，36(23)：72–75.
[8] 倪巍伟，陆介平，陈耿，等. 基于 k 均值分区的数据流高效密度聚类算法[J]. 小型微型计算机系统，2007，28(1)：83–87.
[9] Aggarwal C, Han J, Wang J, et al. A framework for clustering evolving data streams[C]// Proceedings of the 29th VLDB Conference. Berlin, Germany: VLDB Endowment Inc., 2003: 81−92.
[10] 单世民. 基于网格和密度的数据流聚类方法研究[D]. 大连：大连理工大学，2006.
[11] Aggarwal C C, Han J, Wang J, et al. A framework for projected clustering of high dimensional data streams[C]// Proceedings of the 32th VLDB Conference. Toronto, Canada: VLDB Endowment Inc. , 2004: 852–863.
[12] 颜晓龙，沈鸿. 一种适用于高维数据流的子空间聚类方法[J]. 计算机应用，2007，27(7)：1680–1710.

（上接第 38 页）

[3] 王子娴. 胶束增强超滤法去除铬酸盐和硝酸盐[J]. 水处理技术，2005，31(4)：83–84.
[4] 陈蕊贞，任铮伟，冯伟民，等. 乳状型液膜除铬的研究[J]. 华东理工大学学报，1997，23(4)：388–392.
[5] Buerge I J, Hug S J. Kinetics and pH dependence of chromium(VI) reduction by iron(II)[J]. Environmental Science and Technology, 1997, 31(5): 1426–1432.
[6] 俞从正，丁绍兰，孙根行. 皮革分析检验技术[M]. 北京：化学工业出版社，2005：35.
[7] 王伟，丁志农，许佩瑶，等. 几种无机高分子絮凝剂处理制革废水的试验研究[J]. 中国皮革，2008，37(1)：24–27.
[8] 崔淑兰，鞠晓明. 制革铬鞣废水中铬(III)的治理和回收利用[J]. 烟台师范学院学报：自然科学版，1998，14(1)：58–61.
[9] 刘存海. 复合絮凝剂的选配及其在处理铬鞣废水中的应用[J]. 中国皮革，2003，32(5)：28–30.
[10] 赵彤昕，刘金博，赵克军. 混凝沉淀用于铬鞣废液处理方法的探讨[J]. 干旱环境监测，2006，20(2)：119–120.

（上接第 58 页）

参考文献：

[1] 韩渝京，曹勇杰. 首钢 3 号高炉余压发电设计特点与生产实践[J]. 冶金动力，2007(3)：47–49.
[2] 丁洪起，于淑华. S7-400 PLC 在高炉热风炉控制系统的应用[J]. 自动化应用，2010(6)：23–25.
[3] 王志海，冯永海，张娟，等. 基于 S7–400PLC 的煤化工水处理控制系统[J]. 可编程控制器与工厂自动化，2010(10)：113–115.
[4] 李士勇. 模糊控制·神经控制和智能控制论[M]. 哈尔滨：哈尔滨工业大学出版社，2006.
[5] 丁肇红. 液位模糊控制系统设计的应用[J]. 自动化仪表，2005(12)：39–41.