模糊C均值聚类算法在Web使用挖掘上的应用研究

吴 瑛，王秋生
（北京航空航天大学 自动化科学与电气工程学院，北京 100083）

摘 要：Web日志中含有大量的用户浏览信息，从中将相似用户及相关页面进行聚类是建立自适应网站的必要前提。通过基本的预处理，实现了日志的数据净化、用户识别会话识别及数据规约，形成了用户访问页面的序列数据库，同时通过离散化技术计算出用户访问页面频度。在这些数据准备工作的基础上，构造了用户－页面关联矩阵，作为改进的模糊C均值聚类算法的输入，实现了相似用户及相关页面的聚类。实验表明改进的FCM算法的有效性。

关键词：模糊C均值聚类；Web日志预处理；关联矩阵；用户聚类；页面聚类
中图分类号：TP311 **文献标识码**：A **文章编号**：1673－629X(2008)06－0032－04

Research on Application of Fuzzy C－Means Algorithm in Web Usage Mining

WU Ying, WANG Qiu-sheng
(Institute of Automation Science and Electricity Engineering, Beihang University, Beijing 100083, China)

Abstract: Web logs contain a lot of user browsing information. Clustering of similar customers and relative pages is necessary for creating adaptive web sites. Implements the web log's cleaning, user－recognizing, session－recognizing and data convention by means of preprocessing technology. Then a user－page sequence database can be achieved. Simultaneously, the frequency of the user's visit is added to the database. After all these preparation work, can get the associated matrix which is also the input of the improved fuzzy c－means algorithm. Finally realize the clustering of similar customers and relative pages. The result of experiment shows the validity of the algorithm.

Key words: fuzzy c－means algorithm; Web log's data preparation; associated matrix; customer－clustering; page－clustering

0 引 言

Web使用挖掘是通过挖掘Web日志来发现用户访问Web页面的模式。用户访问一个网站时该网站Web服务器会对每次访问进行记录，形成网站的Web日志文件。这些数据可以反映出人们浏览网页的行为模式。目前Web使用挖掘具体算法的研究热点之一是模糊聚类算法，可以用于进行相似用户聚类和相关页面聚类等。传统的聚类分析是一种硬划分，它把每个待识别的对象严格地划分到某个类中，具有非此即彼的性质，因此这种分类的类别界限是分明的。而实际上大多数对象并没有严格的属性，它们在形态和类属方面存在着中介性，适合进行软划分。Zadeh提出的模糊集理论为这种软划分提供了有力的分析工具，人们开始用模糊的方法来处理聚类问题。在模糊聚类算法中又以模糊c均值聚类算法[1](Fuzzy c－means，简称FCM)应用最为广泛。

1 FCM算法相关介绍

FCM是用隶属度确定每个数据点属于某个聚类的程度的一种聚类算法。1973年，Bezdek提出了该算法，作为早期HCM方法的一种改进。FCM将n个向量$x_i(i=1,2,\cdots,n)$分为c个模糊组，并求每组的聚类中心，使得非相似性指标的价值函数达到最小。与引入模糊划分相适应，隶属矩阵U允许有取值在0,1间的元素。不过，加上归一化规定，一个数据集的隶属度的和总等于1：

$$\sum_{i=1}^{c} u_{ij} = 1, \forall j = 1,\cdots,n \tag{1}$$

那么，FCM的价值函数就是

$$J(U, c_1, \cdots, c_c) = \sum_{i=1}^{c} J_i = \sum_{i=1}^{c}\sum_{j}^{n} u_{ij}^m d_{ij}^2 \tag{2}$$

这里$u_{ij} \in [0,1]$；c_i为模糊组I的聚类中心，$d_{ij} = \| c_i - x_j \|$为第i个聚类中心与第j个数据点间的欧几里德距离；且$m \in [1,\infty)$是一个加权指数。

构造如下新的目标函数，可求得使(2)式达到最

收稿日期：2007－09－12
作者简介：吴 瑛(1983－)，女，安徽人，硕士研究生，研究方向为数据仓库及数据挖掘；王秋生，副教授，研究方向为信号处理。

小值的必要条件:

$$\bar{J}(\boldsymbol{U},c_1,\cdots,c_c,\lambda_1,\cdots,\lambda_n) = J(\boldsymbol{U},c_1,\cdots,c_c) + \sum_{j=1}^{n}\lambda_j\left(\sum_{i=1}^{c}u_{ij}-1\right) = \sum_{i=1}^{c}\sum_{j}^{n}u_{ij}^m d_{ij}^2 + \sum_{j=1}^{n}\lambda_j\left(\sum_{i=1}^{c}u_{ij}-1\right) \quad (3)$$

这里 $\lambda_j(j=1,\cdots,n)$ 是式(1)的 n 个约束式的拉格朗日乘子。对所有输入参量求导,使式(2)达到最小的必要条件为:

$$c_i = \frac{\sum_{j=1}^{n}u_{ij}^m x_j}{\sum_{j=1}^{n}u_{ij}^m} \quad (4)$$

和

$$u_{ij} = \frac{1}{\sum_{k=1}^{c}\left(\frac{d_{ij}}{d_{kj}}\right)^{2/(m-1)}} \quad (5)$$

由上述两个必要条件,FCM算法是简单的迭代过程。在批处理方式运行时,FCM用下列步骤确定聚类中心 c_i 和隶属矩阵 $\boldsymbol{U}$:

(1)用值在0,1间的随机数初始化隶属矩阵 $\boldsymbol{U}$,使其满足式(1)中的约束条件。

(2)用式(4)计算 c 个聚类中心 $c_i, i=1,\cdots,c$。

(3)根据式(2)计算价值函数。如果它小于某个确定的阈值,或它相对上次价值函数值的改变量小于某个阈值,则算法停止。

(4)用(5)计算新的 $\boldsymbol{U}$ 矩阵。返回步骤(2)。

上述算法也可以先初始化聚类中心,然后再执行迭代过程。由于不能确保FCM收敛于一个最优解,算法的性能依赖于初始聚类中心。因此,要么用另外的快速算法确定初始聚类中心,要么每次用不同的初始聚类中心启动该算法,多次运行FCM。

2　Web日志挖掘预处理

2.1　Web日志格式

服务器日志的格式根据Web服务的应用及安装时的选项而有所不同,但大多数Web日志一般用两种格式存储[2]:普通日志文件格式(CLF)及扩展日志文件格式(ECLF)。普通日志文件存储的是客户端IP、用户名、状态、服务器名、协议版本等客户连接的物理信息。扩展日志文件格式主要支持关于日志文件元信息的指令,如会话监控开始时间和浏览器类型等。下面是一条典型日志记录及其域名解释:

2004-12-13 0:00:45 172.16.96.22-211.66.184.35

80 GET /~janyst/chat/chatUsers.php - 200 Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1)

* Date与Time:记录了访问网站的访问日期与时间;

* c-ip:IP地址字段记录了发出请求的客户端的IP地址;

* userName:此域保存用户提供的登录名,大多情况该域信息为"-";

* s_ip与s_port:记录客户端访问网站的IP地址与服务器的端口号;

* cs_method:访问者的请求命令,常见的方法有三种分别是GET、POST和HEAD;

* cs_uri_stem:记录网站被访问的资源;

* cs_uri_query:客户所执行的查询;

* sc_status:服务器返回的状态代码;

* userAgent:记录产生请求的用户代理的信息,包括产品代号、用户浏览器和用户操作系统的类型。

2.2　Web日志预处理

预处理过程主要包括四个阶段:数据净化、用户识别、会话识别及数据归约。

数据净化指删除Web日志中的无效数据。文中实验中无效数据主要包括以下几种情况:cs_uri_stem扩展名为jpg、gif等。

用户识别是将用户和请求的页面相关联的过程。用户识别方法主要有三种[3~5],分别为基于cookie的技术、基于IP地址以及基于网络拓扑结构的路径分析。实验用到的日志不仅包括c_ip且包括userAgent域,所以采用基于IP地址和用户浏览器的方法进行用户识别,即不同的IP地址和用户浏览器代表不同的用户,这样可以更加精确地识别用户。

会话(Session)是指用户在一次访问网站期间从进入网站到离开网站所进行的一系列活动。用户会话[6]是一个二元组<userID,pageID>,其中userID是会话标识,PageID是用户在该某时间段内请求Web页面的集合。要构造一个会话就是将每个用户的活动日志按照某种方法映射到会话中的过程。识别一个会话的方法主要有两种:基于时间的启发式方法和基于引用的启发式方法。文中采用基于时间的启发式方法进行会话识别,即同一用户依次发出相邻的页面请求之间的时间间隔如不超过时间阈值,那么这两个页面请求属于同一个会话。文中将时间阈值设定为20分钟,实验证明效果良好。

数据归约是指根据识别得到的不同用户、不同会话与页面进行规约,把同一用户在相同会话中访问过的页面放置在一起,作为该用户的访问记录。

3 FCM 在日志挖掘中的应用

3.1 数据准备工作

实验选取某大学校园网的一段 Web 日志，将通过上述预处理后，可形成表 1，表中各字段属性分别为：userID 表示用户识别后的用户标识号；sessionID 表示会话识别后的会话标识号；pageID 代表各个页面标识号；timeID 代表访问页面时刻。

表 1 用户－会话－页面表

文件(F) 窗口(W) 帮助(H)

userID	sessionID	pageID	timeID
1	1	1	2004-12-13 00:00:45
1	1	2	2004-12-13 00:00:45
1	127	49	2004-12-13 00:41:16
1	127	24	2004-12-13 00:41:48
1	127	15	2004-12-13 00:43:26
1	127	14	2004-12-13 00:43:26
1	127	23	2004-12-13 00:43:40
2	2	3	2004-12-13 00:02:26
2	2	2	2004-12-13 00:02:29
2	2	4	2004-12-13 00:02:37
2	2	5	2004-12-13 00:02:45
2	2	6	2004-12-13 00:03:18
2	124	3	2004-12-13 00:40:23
2	124	2	2004-12-13 00:40:27
3	3	3	2004-12-13 00:02:47
4	4	3	2004-12-13 00:06:50
4	4	2	2004-12-13 00:06:52
5	5	3	2004-12-13 00:07:12
5	5	2	2004-12-13 00:07:13
5	230	3	2004-12-13 01:07:23
5	230	2	2004-12-13 01:07:25
6	6	3	2004-12-13 00:07:29
6	6	2	2004-12-13 00:07:30
7	7	7	2004-12-13 00:07:46
8	8	3	2004-12-13 00:09:08
8	8	2	2004-12-13 00:09:09
8	8	8	2004-12-13 00:09:09
8	8	10	2004-12-13 00:10:20
9	9	9	2004-12-13 00:09:51

根据表 1 中的 timeID 值计算出每个用户在每次会话中访问特定页面的停留时间 timestay，再采用如表 2 的离散化技术将其转化为新的属性 visit，表示用户访问特定页面的频度。采用这种浏览时间离散化的表式方法，用户只要访问了页面，即使时间再短也有离散化时间(值为 -1)；用户浏览页面时间即使很长，也有离散化时间(值为 19)，这样在进行用户聚类及页面聚类时可以更准确地进行相似性度量。如表 1 经过离散化过程后可以转化为表 3。

3.2 用户聚类

相似用户聚类主要是把用户划分为若干组，具有相似浏览模式的用户分在一组，它一般用于电子商务中，为用户提供个性化服务。假设某网站有 n 个页面，用 $P=\{P_1,P_2,\cdots,P_n\}$ 表示。在某时间段，有 m 个用户访问该网站，用 $U=\{U_1,U_2,U_3,\cdots,U_m\}$ 来表示。实现用户聚类步骤如下：根据预处理后的日志文件，计算每个用户 $U_i(i=1,2,\cdots,m)$ 在该时间段访问各页面 $P_j(j=1,2,\cdots,n)$ 的频度，用 $C(U_i,P_j)$ 表示；从而得到一个 $m*n$ 的样本矩阵 $\mathbf{A}$；运行 FCM 算法对输入样本进行分析。

表 2 访问时间离散化

离散化值(根据数据值离散化)	访问情况(单位:s)
-3	访问该页面以后彻底离开
-2	访问页面后暂时离开
-1	访问一次离开
1	0s<=页面访问时间<20s
2	20s<=页面访问时间<40s
3	40s<=页面访问时间<60s
4	60s<=页面访问时间<80s
5	80s<=页面访问时间<100s
6	100s<=页面访问时间<200s
7	200s<=页面访问时间<300s
8	300s<=页面访问时间<400s
9	400s<=页面访问时间<500s
10	500s<=页面访问时间<600s
11	600s<=页面访问时间<700s
12	700s<=页面访问时间<800s
13	800s<=页面访问时间<900s
14	900s<=页面访问时间<1000s
15	1000s<=页面访问时间<1500s
16	1500s<=页面访问时间<2000s
17	2000s<=页面访问时间<2500s
18	2500s<=页面访问时间<3000s
19	3000s<=页面访问时间

表 3 用户－页面－频度表

文件(F) 窗口(W) 帮助(H)

userID	sessionID	pageID	visit
1	1	1	-1
1	1	2	-2
1	127	49	2
1	127	24	5
1	127	15	-1
1	127	14	1
1	127	23	-3
2	2	3	1
2	2	2	1
2	2	4	1
2	2	5	2
2	2	6	-2
2	124	3	1
2	124	2	-3
3	3	3	-3
4	4	3	1
4	4	2	-3
5	5	3	1
5	5	2	-2
5	230	3	1
5	230	2	-3
6	6	3	1
6	6	2	-3
7	7	7	-3
8	8	3	1
8	8	2	-1
8	8	8	4
8	8	10	-3
9	9	9	-3

3.3 页面聚类

相关页面聚类是找出具有相关内容的网页组，这对网上搜索引擎和调整页面结构等应用很有意义。页面聚类和用户聚类类似。实现用户聚类步骤如下：根据预处理后的日志文件，计算每一页面 $P_i(i=1,2,$

…,n)在该时间段被各用户 $U_j(j=1,2,\cdots,m)$ 访问的频度,用 $C(P_i,U_j)$ 表示;从而得到一个 $n*m$ 的样本矩阵 $\boldsymbol{B}$;运行 FCM 算法对输入样本进行分析。

3.4 算法描述

算法:改进后 FCM (Matlab)

输入:原始关联矩阵 $\boldsymbol{X}$,样本的个数 sample_num,每个样本的维数 dimension,要聚类的类别数 category,最大迭代次数 maxcycle,参数 m,误差限 limit;

输出:相似用户/相关页面集合

```
%随机初始化聚类中心
Cmat=fcm_dataInitC(X, category);
%计算各样本点和各聚类中心的距离矩阵
Dmat=fcm_calcD(X, Cmat);
%计算初始的模糊相似矩阵
Umat=fcm_calcU(m, Dmat);
%开始迭代
U=Umat; int cycle_num=1; boolean flag=0;
while (cycle_num<maxcycle & flag~=1)
%根据 X、U 重新计算 Cmat
Cmat=fcm_calcC(Xdat, category, m, U);
%根据新的 Cmat 计算 Dmat
Dmat=fcm_calcD(X, Cmat);
%更新 Umat
Umat=fcm_calcU(m, Dmat);
%计算 U 和 Umat 之间的差距
U=U-Umat;
%计算 U 矩阵范数 Frobenius 范数
Fro_norm=norm(U,'fro');
if (Fro_norm<limit)
flag=1;
else
U=Umat;
end
cycle_num=cycle_num +1;
end
%计算最终聚类中心
fcm_evaluatecenter(Umat, m, Dmat);
return
```

3.5 实验结果

按照上文中所述步骤可得到样本矩阵 $\boldsymbol{A}$、$\boldsymbol{B}$ 作为算法输入,再运用所描述的改进后 FCM 算法对 Web 日志进行聚类分析。在此,以日志中的一个子集为例,具体数据如表1。这里列举了由9个用户组成的用户集合 $\boldsymbol{U}=\{U_1,U_2,U_3,\cdots,U_9\}$ 以及由 15 个不同 cs_uri_stem 属性的页面所组成的页面集合 $P=\{P_1,P_2,P_3,P_4,P_5,P_6,P_7,P_8,P_9,P_{10},P_{14},P_{15},P_{23},P_{24},P_{49}\}$,由此可得到一个 9×15 的用户矩阵如表4。

表4　用户矩阵

	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9	P_{10}	P_{14}	P_{15}	P_{23}	P_{24}	P_{49}
U_1	-1	-2	0	0	0	0	0	0	0	0	1	-1	-3	5	2
U_2	0	-2	2	1	2	-2	0	0	0	0	0	0	0	0	0
U_3	0	0	-3	0	0	0	0	0	0	0	0	0	0	0	0
U_4	0	-3	1	0	0	0	0	0	0	0	0	0	0	0	0
U_5	0	-5	2	0	0	0	0	0	0	0	0	0	0	0	0
U_6	0	-3	1	0	0	0	0	0	0	0	0	0	0	0	0
U_7	0	0	0	0	0	0	-3	0	0	0	0	0	0	0	0
U_8	0	-1	1	0	0	0	0	4	0	-3	0	0	0	0	0
U_9	0	0	0	0	0	0	0	0	-3	0	0	0	0	0	0

实验中选用参数 category = 3、maxcycle = 500、m = 2、limit = 0.1。根据改进后的FCM算法,可以得到用户聚类后的结果,将此9个用户分为3类:$\{U_1,U_2,U_8\}\{U_4,U_5,U_6\}\{U_3,U_7,U_9\}$。

同理,可以得到一个 15×9 的页面矩阵,即上述用户矩阵的转置矩阵 $\boldsymbol{B}=\boldsymbol{A}^{\mathrm{T}}$。依然选用上述参数,可以将这 15 个页面分为三类:$\{P_1,P_3,P_4,P_5,P_8,P_{15},P_{23}\}\{P_2,P_6,P_7,P_9,P_{10}\}\{P_{14},P_{24},P_{49}\}$。

结合页面的内容,可发现聚类结果与预期相符合,证明了 FCM 算法在 Web 日志挖掘中的有效性。

4 结束语

通过挖掘 Web 日志中隐含的知识将相似用户及相关页面进行聚类是建立自适应网站的必要前提。传统的聚类分析是一种硬划分,在 Web 日志挖掘实际应用中有其局限性,因此根据 Web 结构特点和模糊聚类算法思想实现了模糊 c 均值算法在 Web 日志挖掘中的应用,对原始的 Web 日志进行了用户及页面上的聚类,该算法容易理解,可扩展性良好。实验结果表明算法有效的同时,也提出了一个重要问题,即如何评价结果的准确度,这是下一步的工作重点。

参考文献:

[1] Bezdek J C. Fuzzy Mathematics in Pattern Classification[D]. Ithaca: Applied Math. Center, Cornell University, 1973.

[2] Sweiger M, Madsen M R, Langston J. 点击流数据仓库[M]. 陆昌辉译. 北京:电子工业出版社,2004.

[3] Pitkow J. In Search of Reliable Usage Data on the WWW[C]// In: sixth International World Wide Web Conference. Santa Clara, CA: [s.n.], 1997:451-463.

[4] Cooley R, Mobasher B, Srivasta J. Data Prepatation for mining world wide Web browsing patterns[J]. Journal of Knowledge an Information System, 1999,1(1):5-32.

[5] Pirolli P, Pitkow J, Rao R. Silk from a Sow's Ear: Extracting Usable Structure from the Web[C]// In: Proceedings of CHI'96. Vancouver BC: ACM Press, 1996:118-125.

[6] 杜家强,韩其睿,王　科,等. Web 日志中用户频繁路径快速挖掘算法[J]. 计算机工程与应用,2005(22):79-83.