doi:10.3969/j.issn.1000-1247.2017.07.007

基于点播行为特征指数的IPTV用户精准分群算法

汪敏娟[1,2]　吕　超[1,2]
1.中国电信股份有限公司智慧家庭运营中心
2.江苏省公用信息有限公司

摘要　提出一种提取IPTV用户有效点播行为的算法：首先基于有限状态机对用户点播行为进行实时监测，将符合实际点播行为规则的数据计入用户行为库；在用户行为库的基础上通过用户点播行为特征指数分析将无效点播行为进行过滤，得到精准描述用户点播行为特征的向量，将独立的IPTV用户抽象为以点播行为特征描述的向量；基于精确点播行为向量对用户进行相似度分析、聚类，实现对用户的精准分群；最后通过大量的实际IPTV运营数据对算法模型进行验证分析。

关键词　有限状态机　实际点播行为　点播行为特征指数　基于余弦定理的用户分群

1 引言

用户的点播行为是指用户对IPTV论域内海量但有限的内容进行点播。在实际业务运营中，IPTV内容都分配相应的内容标签，通过分析独立IPTV用户对内容的点播及驻留情况，即可有效地将独立用户抽象为对论域内内容标签偏好量化描述的向量。传统的算法中对用户与标签的偏好量化描述提供了较为有效的计算模型，但在实际运营过程中，由于IPTV的媒体属性，用户使用IPTV时的点播行为是自然而然甚至无意识的。故在基于用户内容点播行为抽象用户对内容标签的偏好时需要对用户的点播行为进行有效过滤，将用户非主观、随机点播的行为进行过滤，这样得到的规则较为贴近实际运营情况，而目前从上述角度进行研究的工作较少。

为此，从用户的点播行为入手，基于点播行为状态机来获取用户的实际点播行为，即符合预设标准的点播行为；并在用户实际点播行为的基础上计算内容标签点播时长对点播行为特征描述“影响程度”的量化值，过滤用户的无意识点播行为，从而将IPTV用户抽象为对内容标签的有效偏好向量，基于该向量对用户进行精准分群。下面通过算法在实际运营应用数据的实验进行分析验证。

2 IPTV用户实际点播行为采集

IPTV用户精确分群算法模型是以用户点播行为数据为基础进行噪音过滤和分群聚类的数学模型。首先通过点播行为判定状态机对用户在IPTV使用过程中的实际点播行为进行采集和分析，将属于用户的实际点播行为进行记录。

IPTV由EPG、Content（可点播内容）构成。EPG根据面向用户展示情景的不同分为Portal（首页/导航）、Column/Sub Column（栏目和子栏目）等状态，用户从Portal开始通过多次点击操作至Content。用户进入可点播内容状态后，根据用户的实际驻留时间产生实际点播行为。为有效获取到用户的实际点播行为，基于IPTV的组成和用户点播行为特征，定义用户点播行为状态机。

定义1.用户点播行为状态机$FSM_{UC}=\{I,P,f,\delta\}$。$I$为用户点击操作；$P$为非空状态集，在IPTV论域内定义为面向用户展示的各EPG状态和最终状态；f为最终状态，在IPTV论域内定义为可点播内容状态，$f\in P$；δ为状态转移函数，$\delta:P\times I:P$，即用户通过点击操作在不同的EPG状态集P中进行转换直至到达最终状态f。点播行为状态机描述了从开机开始，直至退出服务期间的点播状态变化情况。

设IPTV可点播内容集$C=\{c_1,c_2,\cdots,c_n\}$，IPTV可点播内容集C记录了IPTV论域内可被用户点播的全量内容。设IPTV内容标签集$T=\{t_1,t_2,\cdots,t_m\}$，其中$\forall t_i\in T$表示IPTV论域内IPTV内容的特征属性标记。在IPTV论域内，任意IPTV内容均对应一个内容标签集合，即$\forall c_i\in C$，都有$\exists T'\in T$且$T'\neq\phi$，使T'中的内容标签描述内容c_i的内容特征。

定义2.用户实际点播行为向量$RC_i=\{a_{i1},a_{i2},\cdots,a_{im}\}$，$a_{ij}$表示用户$u_i$在内容标签$t_j\in T$被实际点播的时长。

为加强用户行为特征描述的准确性，用户点播行为状态机可有效地对用户点播行为进行跟踪：从初始状态开始，直至进入可点播内容状态f，有限状态机可实时记录用户状态

的变化情况，当用户进入可点播内容状态f后，要求状态机记录用户停留在状态f的时长；当符合预设阈值时，即认定用户当前“进入可点播内容”状态有效，在用户实际点播行为集中将当前可点播内容对应的内容标签实际时长进行相应增加。

将用户的收视点播行为抽象为有限状态机后，有限状态机作为成熟的数学模型，利用现有的各类处理工具即可便捷地对用户有效点播行为进行实时采集和处理。对符合有效点播阈值的点播行为，在该用户对应的实际点播行为向量中，为点播内容标签增加相应的实际点播时长。

3 IPTV有效点播行为过滤

实际点播行为向量记录了IPTV用户在收视过程中达到预设标准的点播行为。通过对用户实际点播行为的采集将用户抽象为对IPTV论域内各内容标签的实际点播行为向量。向量中每一个元素都是对用户在IPTV论域内的内容标签的实际使用情况描述。理论上用户的行为特征可以通过用户对内容标签使用的时长计算得到，即计算用户实际点播行为向量中各标签实际使用时长与总时长的比值，当比值超过一定阈值时可判定用户对该类标签具备行为偏好。

内容标签实际点播频率$CF_{ij}=a_{ij}/\sum_{k=1}^{m}a_{km}$描述用户$u_i$对内容标签$t_j$的实际点播频率。其中$a_{ij}$表示用户$u_i$在内容标签$t_j\in T$被实际点播时长；$\sum_{k=1}^{m}a_{km}$表示用户$u_i$实际点播行为向量各标签的实际点播总时长。

在实际运营过程中发现，由于电视屏业务“无意识操作”的特征，用户的实际点播行为记录中仍存在较多的噪音数据。虽然部分内容标签被点播时长较多，达到实际点播行为采集阈值，但此标签可能较多次播放为用户的“无意识操作”，无法准确反映用户行为特征。

在用户实际点播行为向量中，不同的内容标签所包含的用户点播行为特征信息量各不相同，仅通过内容标签实际点播频率无法准确地描述用户偏好信息。因此要对内容标签对用户行为特征的“影响程度”进行量化分析，并将用户在各内容标签上的点播时长、各内容标签对用户的“影响程度”结合起来描述点播行为，过滤用户点播行为中的噪音数据，从而能够获取到对用户分类起方向性作用的分类标准。

在经典信息论中定义了“逆频率指数”，通过计算信息片段在论域知识内的出现频率来判定信息片段对知识整体的影响程度。“逆频率指数”越高，表明信息片段在越少量的知识中出现，则该信息片段对知识越具有描述能力。在IPTV论域内，实际点播时长对用户行为特征的描述与之相反，即内容标签点播时长信息在用户整体行为中出现的次数越少，则表明其为“无意识”操作的可能性越大。基于经典“逆频率指数”定义，结合IPTV论域内实际点播时长与行为特征的关系，下面定义“点播行为特征指数”来描述内容标签播放时长对内容点播行为特征描述的“影响程度”值。该权重值量化的描述内容标签t_j对用户u_i的点播行为特征的表述程度，从而确保实际点播行为向量中各元素更为真实地描述用户点播行为特征。

点播特征指数$ICF_{ij}=\lg(a_{ij}/\sum_{k=1}^{m}a_{km})$描述内容标签$t_j$对用户$u_i$实际点播行为特征的量化描述能力。通过点播特征指数对用户实际点播行为向量进行加权，将用户的实际点播行为向量抽象为对IPTV论域内全量内容标签的行为特征描述。

点播行为特征向量$RRC_i=\{r_{i1},r_{i2},\cdots,r_{im}\}$，$r_{ij}=CF_{ij}\times ICF_{ij}$，其中$CF_{ij}$表示用户$u_i$对内容标签$t_j$的实际点播频率，$ICF_{ij}$表示内容标签$t_j$对用户$u_i$实际点播行为特征的量化描述能力。通过点播特征指数，可将用户实际点播行为向量进一步抽象，从而将用户抽象为对IPTV论域内各内容标签点播偏好量化描述的点播行为特征向量。

利用点播行为特征向量，可精确地对用户进行分群。两个独立的IPTV用户u_i、u_j相似度可通过用户实际点播行为向量之间的夹角余弦来判定，即：

IPTV用户相似度$\mathrm{sim}(u_i,u_j)=\sum_{k=1}^{m}ri_k\times r_{jk}/\sqrt{\sum_{k=1}^{m}r_{ik}^2}\times\sqrt{\sum_{k=1}^{m}r_{jk}^2}$

$\mathrm{sim}(u_i,u_j)$基于点播行为特征对IPTV用户u_i、u_j的“相似程度”进行量化描述。r_{ik}、r_{jk}为用户u_i、u_j对应的加权实际点播行为向量元素。当$\mathrm{sim}(u_i,u_j)=1$时，表示IPTV用户u_i、u_j完全相同，当$\mathrm{sim}(u_i,u_j)=0$时，表示用户u_i、u_j完全不同。对IPTV论域内用户计算其相似度，直至一组用户的相似度收敛到预设的阈值，完成IPTV论域内用户的分组。通过上述分群模型分析，各用户分组对IPTV论域内的内容标签具有相似的偏好。

4 实验分析

对N市IPTV用户基于点播行为特征指数进行分群实验。首先基于N市IPTV可点播内容集建立IPTV内容标签集。基于N市用户1～11月的点播行为历史记录，通过有限状态机建立N市用户实际点播行为向量。基于实际点播行为向量计算点播行为特征向量，利用点播行为特征向量对用户进行分群，并对各分群中用户在12月的实际点播行为进行比对，以验证分组的有效性。

按照提出的用户基本分类算法，将N市用户分成了10个用户基本分类后停止分类，从表1中可以看到，通过11个月的历史数据分析形成的用户分群对于IPTV内容标签偏好程度各不相同。

对各用户分类群在12月的实际点播行为进行汇总，筛选出各用户分群在12月的有效点播行为，将各分群的标签偏好

表1　IPTV内容标签集

用户分群编号	各分群内容标签	各分群用户数（万户）
1	文艺、爱情、家庭	26.4
2	惊悚、动作、科幻、武侠	22
3	纪实、美食、人文	5.7
4	偶像、综艺、爱情	30.7
5	新闻、人文、纪实	12
6	喜剧、家庭、生活	6.9
7	音乐、偶像、综艺	4.2
8	体育、竞技、游戏	5.5
9	法制、家庭、新闻	4.4
10	养生、家庭、美食	9

表2　预测规则与实际规则对比

分群编号	各分群内容标签	12月有效标签对应用户数（万户）	匹配度
1	文艺、爱情、家庭	23.9	90.5%
2	惊悚、动作、科幻、武侠	21.3	96.8%
3	纪实、美食、人文	5.5	96.5%
4	偶像、综艺、爱情	27.8	90.6%
5	新闻、人文、纪实	11.2	93.3%
6	喜剧、家庭、生活	6.5	94.2%
7	音乐、偶像、综艺	3.9	92.9%
8	体育、竞技、游戏	5.1	92.7%
9	法制、家庭、新闻	4.1	93.2%
10	养生、家庭、美食	8.5	94.4%

与有效点播行为对应的标签偏好进行对比，用于验证预测规则的准确性。对比情况见表2。

从表2可以看到，各用户分群对内容标签的偏好程度和12月实际点播行为的误差率都在10%以内，用户分群较为准确。对偏差较大的用户分群1、4，进行抽样调查，分群中的用户在日常IPTV点播时随意性较大，虽然有一定的内容类型偏好，但并未有如其他分群中用户“对分类有非常明确的方向性导向”的行为特征。算法通过“有效点播行为”过滤得到的标签，在进行用户调查时，基本得到用户“有一定点播偏好”的正面反馈。

5 结束语

提出一种符合IPTV用户点播行为的分群算法，结合实际运营经验将IPTV用户的点播行为进行有效过滤。首先通过有限状态机对非有效点播行为进行过滤，得到用户的实际点播行为向量。在实际点播行为向量的基础上，通过一段时期用户的点播行为分析，过滤用户在点播过程中非主观、无意识的点播行为数据，进一步降低用户行为向量中的数据噪音，使向量能够量化描述用户对IPTV论域内容标签的偏好程度，从而得到点播行为特征向量。基于独立IPTV用户的点播行为特征向量，利用余弦定理来计算用户间相似程度的量化值，从而进行有效、准确地用户分群。通过对实际运营数据的分析，提出的算法明显地提升了IPTV用户分群的准确性，降低了IPTV用户分类的计算开销。

在后续的工作中，需对当前的过滤算法进一步进行精确，对IPTV用户因媒体特征产生的无意识点播行为进行深入过滤，使算法获取的用户点播行为能够充分地描述用户的实际偏好，并优化对IPTV用户分类算法和计算规则的约定等研究工作，进一步通过大规模数据统计规律优化对用户点播行为采集的估算方法。

参考文献

[1] MW Kim,WM Song,SY Song,EJ Kim.Efficeint Collaborative Recommendation with Users Clustered for IPTV Services.Springer Berlin Heidelberg[J].2012,310

[2] 王泽，屈海伟.大数据分析在IPTV维护中的应用[J].电信技术,2014,1(10)

[3] 苏军根，杨柳，武娟.IPTV用户数据挖掘建模体系研究[J].电信技术,2014,1(2)

[4] Jim,Analysis of the IPTV Increment Service Decvelopment Strategy[C].Science&Technology Innovation Herald,2014

[5] H Park,K han,J Yang,JK Choi.Enhanced Metadata Creation and Utilization for Personalized IPTV Service[C].International Conference on Information Science and Applications,2017

[6] LI Zhihua,Y Sun,LI Zuopeng. Design and implementation of IPTV terminal network management authentication System[J].Video Engineering,2017

[7] J Zhao,K Yang,X wei,Y Ding,etc.A Heuristic Clustering—Based Task Deployment Approach for Load Balancing Using Bayes Theorem in Cloud Environment[J].IEEE Transactions on Parallel and Distributed systems,2016,27(2)

如对本文内容有任何观点或评论，请发E-mail至ttm@bjxintong.com.cn。

作者简介

汪敏娟

高级工程师，现就职于中国电信股份有限公司智慧家庭运营中心、江苏省公用信息有限公司。主要研究方向包括：IPTV运营规律、视频承载网络规划、知识学习。

吕超

工程师，现就职于中国电信股份有限公司智慧家庭运营中心、江苏省公用信息有限公司。主要从事大数据平台设计、数据运营与维护工作。