基于用户行为聚类的人物角色量化模型创建实证研究

孙敏杰[1,2] 吴振新[2]

[1](中国科学院国家科学图书馆 北京 100190)

[2](中国科学院研究生院 北京 100049)

【摘要】根据人机交互设计中人物角色的用户建模思想,在机构仓储系统的应用环境下,通过对用户行为日志的分析,采用 K-means 聚类方法识别用户行为模式,并据此划分主要用户群体类型,创建机构仓储系统的人物角色-行为特征矩阵量化模型。

【关键词】人物角色 用户行为 用户建模 聚类 机构仓储系统

【分类号】G250

User Behavior Clustering for Creation of Personas

Sun Minjie[1,2] Wu Zhenxin[2]

[1](National Science Library, Chinese Academy of Sciences, Beijing 100190, China)

[2](Graduate University of Chinese Academy of Sciences, Beijing 100049, China)

【Abstract】Based on the personas of user modeling in human-computer interaction design, through the analysis of user behavior logs in institutional repository, the authors use K-means clustering method to identify user behavior patterns, classify users group, and create personas-feature matrix quantitative models for institutional repository.

【Keywords】Personas User behavior User model Cluster Institutional repository

1 引 言

随着信息技术的飞速发展,用户的信息环境及信息行为在不断变化,用户信息行为的改变和追踪成为众多领域的关注热点。2010 年 5 月,OCLC 与 JISC 联合发布了图书馆“用户信息探索行为”的系列报告[1],作为 OCLC、RIN 和 JISC 的“用户行为项目”的部分研究成果,旨在感知用户的信息探索行为,以便使图书馆的信息服务和系统能够更好地满足用户需求。

研究表明,同类用户群体往往具有共同的行为模式,行为模式是从大量实际行为中概括出来的,作为行为的理论抽象、基本框架或标准。它反映了用户为完成任务而在行为上表现出的规律性,据此可根据用户已有行为预测其未来可能的行为。本文引入人机交互领域中人物角色的用户建模方法,通过用户行为日志分析,识别用户群体的行为模式,进而区分用户群体的角色类型,以更好地辅助图书馆开展个性化的信息服务。

2 研究背景

作为人物角色方法的创立者,Cooper 定义人物角色是“精确描述用户和用户想要完成的事”[2]。其典型特征是:

收稿日期:2010-06-18
收修改稿日期:2010-11-20

(1)人物角色是真实用户的假想原型;

(2)人物角色由用户的目标来定义,同一角色的用户往往具有相似的特征、共同的行为模式;

(3)人物角色特征的描述要详尽、具体、精确。

他认为人物角色关键在于如何确定角色描述和怎样使用角色描述,即人物角色的创建与应用。其过程分为4个阶段:

(1)用户信息获取:用户特征因素是角色描述的内容来源,通常包括人物目标、人口统计学、行为方式、场景环境、心理与价值取向等;获取途径为用户访谈、直接观察、问卷调查和日志采集。

(2)行为模式识别:决定角色的类别,识别方法有专家经验法、统计分析法、聚类。

(3)人物角色表形成:采用文字描述的方式描述角色及其典型特征。

(4)在交互设计中应用:通过主、次角色的区分,帮助设计师定位主要角色的需求。

可以看出,人物角色方法所生成的角色表目前是一种文字描述型的模型,人们以此为参考来辅助系统的设计,同时其整个生成过程和方法都带有很重的主观的因素。那么能否以客观的方法来产生人物角色、以量化模型的形式来存储和表现角色表,并以动态方式支持数字服务系统的个性化服务?本文以此为切入点,以机构仓储系统(Institutional Repository,IR)为实例,通过解析系统服务日志来获取用户行为模式,划分角色类型,以人物角色的行为特征矩阵来表示IR的人物角色量化模型,初步探讨了人物角色量化模型的创建。

3 人物角色量化模型创建实证研究

在传统人物角色创建原理[3]基础上,本文提出基于用户行为聚类的人物角色量化模型的创建过程,如图1所示:

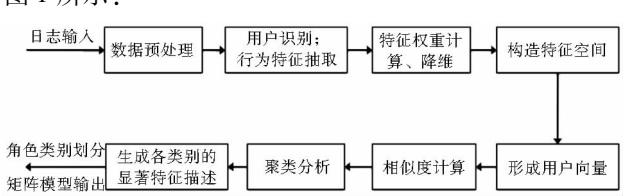

图1 人物角色量化模型创建实验过程

采用向量空间模型①对用户及其行为特征进行表示,然后对用户间进行相似度计算和聚类处理,最终构建人物角色量化模型。

人物角色模型的构建以特征独立性假设为前提,即行为特征在用户中的出现是独立、互不影响的,群体用户的行为特征之间的关联关系所占的比重不是该模型重点考量的对象。

图1中,通过对日志中用户行为特征的抽取,构造用户行为特征空间,并将用户表示为特征空间上的向量;然后通过用户向量间距离的计算判断用户间的相似程度,采用一定的聚类算法识别和汇聚用户类型;最后采用显著行为特征对角色及其行为模式进行描述,并用最终聚类中心矩阵来表示人物角色量化模型。

3.1 准备阶段

本文选择中国科学院国家科学图书馆的机构仓储系统[4](机构环境下的文献资源传播与共享平台)作为实证系统,利用Java开发相关程序完成聚类前的日志清洗和权重计算,同时选择PostgreSQL数据库来存储清洗阶段的数据和聚类分析结果,采用SPSS18.0作为聚类分析及统计工具,完成用户向量间相似度的计算、聚类分析以及聚类结果的统计等工作。

为了更准确地从日志中获得用户行为记录,首先对IR的110多万条日志数据(遵循通用日志格式ECLF标准[5])做预处理:

(1)清理无实际意义日志记录,如:对扩展名为"JPG、GIF、PNG、BMP、ICO"的图片类资源的请求,对CSS样式表的请求和JavaScript脚本请求等。

(2)清理搜索引擎采集日志记录,如Googlebot、MSNBot、Yahoo! Slurp等搜索引擎[6],Bot、Crawler、TestAgent、Spider等关键字。

(3)日志记录分类转存:将日志记录按内容类型逐条拆分并转存到数据库中。

随后,进行用户识别及其行为特征抽取。用户和用户行为特征是本实验最终要构建的人物角色量化模型的组成元素,分别构成了角色矩阵的行向量和列向量。但这两项内容在日志记录中没有确定的标识,需要根据一定的假设和判定规则从日志中识别、抽取出来。

(1)用户识别:假设每一位用户都有固定的上网地址,使用固定的计算机,那么将具有"相同IP地址、相同操作系统、相同Web浏览器"的用户视为同一用户。

① 王斌. 信息检索模型. 中国科学院2008年度信息检索课程讲义.

相对于真实用户的复杂情况来说,该识别规则具有一定的局限性,尤其是当用户使用动态 IP 时,有两类情况会对结果产生比较明显的影响。

①同一用户会因 IP 地址的改变而被识别为不同的用户。这种情况下,由于同一用户在行为模式上的统一性,通常会被聚集到一种角色中,但可能导致该角色包含的用户数量比真实用户数量大。

②具有相同操作系统和 Web 浏览器的不同用户会因为使用过同一 IP 地址而被识别为同一用户。这种情况会对角色的种类划分产生干扰,为尽量减小这类影响,本实验统计了每个用户所包含的行为日志记录数量,如果有超出一般水平的极端情况,则认为未能识别出多个真实用户,将不参加人物角色模型的聚类计算。

(2)用户行为特征抽取:假设每种行为特征都是独立的、唯一的,并由唯一标识符代表该特征出现在 URL 中。

行为特征由关键字或 archivID 在 URL 中标识,每种特征之间有显著的间隔符号。行为特征的分解和抽取过程,如图 2 所示:

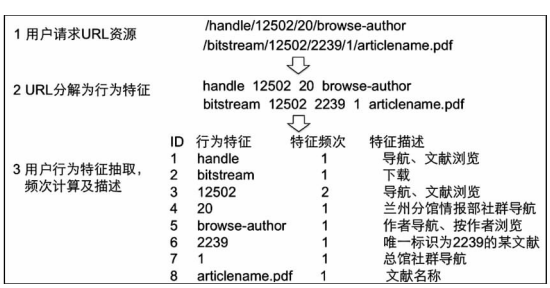

图 2　用户行为特征抽取过程

图 2 中,按"/"等间隔符号分解用户请求的 URL 资源,从中抽取独立的行为特征单元,并描述每个特征在 IR 中所代表的意义。另外,统计每个特征出现的频次,作为后续特征重要程度计算的指标来源之一。

3.2　行为特征空间构造阶段

(1) 用户行为特征权重的计算

采用 TF - IDF 算法[7]来计算用户行为特征权重,评估一种行为特征对于一个用户集合中的某一用户的重要程度,也是该用户有别于其他用户的区分能力。这里,将 TF - IDF 算法思想解释为:行为特征的重要性随着它在某一用户行为中出现的次数成正比增加,但同时会随着它在整个用户集合中出现的频率成反比下降。即如果某个特征在某一用户中出现的频率 TF 高,并且在其他用户中很少出现,则认为该特征对于该用户来说很重要,适合用来做与其他用户间的类别区分。公式如下:

$$a_{ij} = TF_{ij} \times IDF_i \quad (1)$$

$$TF_{ij}\text{归一化}: \frac{TF_i}{\underset{i}{MaxTF_i}} \quad (2)$$

$$IDF_i\ \text{归一化}: \ln(N/DF_i) \quad (3)$$

其中,a_{ij}代表行为特征 $term_i$ 在用户 $user_j$ 中的权重;TF_{ij}归一化后取值为:$term_i$ 在用户 $user_j$ 中出现的次数除以该 user 的 term 总数;IDF_i归一化后取值为:整个用户集合中的用户总数 N 除以出现 $term_i$ 的用户个数,然后再对商值取以 e 为底的对数。

逆文档频率 IDF 对特征区分度会产生比较显著的作用,这一点也在实验中得到了验证。以某一用户的两个行为特征(Term)为例做进一步解释,如表 1 所示:

表 1　IDF 逆文档频率对 Term 区分度的影响实例

ID	Term	TF	DF	MaxTF	N	$a_{ij} = TF_{ij} \times IDF_i$
1294	authorize	79	21	3 509	8 858	(79/3509) × ln(8858/21) = 0.136
1181	simple - search	80	2 235	3 509	8 858	(80/3509) × ln(8858/2235) = 0.031

从表 1 可以看出,两个特征在其他指标都相近的情况下,由于在整个用户集合中出现"authorize"这一行为特征的用户数量为 21,其值明显小于"simple - search"的用户数量 2 235,因此前者的权重要大于后者。对于这一用户来说,他的"authorize"行为特征要比"simple - search"行为特征更显著。

(2) 特征降维及特征空间构造

在文档向量空间模型中,当表示文档的词特征向量维度过高时,不仅会大大增加计算的耗时,也会降低算法的精度,因此会先对特征进行降维处理。Sinka 等认为,当选择 1% 的高频词时,可以将辨识能力较强的词纳入特征空间,同时将辨识能力不强的词排除在外,使得聚类效果最好[8]。

通过对特征频率和用户分布情况的统计,先排除无效高频特征和极低频特征,然后根据帕欧公式[9]在剩下的特征中进一步圈定有效特征的范围,最终将有效特征的频率范围定为 100 - 17 154 次之间,特征种数为 223 种,构建了用户行为特征空间。

3.3　用户向量形成阶段

在向量空间模型中,用户可以看作由多个行为特征所组成的坐标轴上的点,因此用户向量可以由行为

特征及其权重来表示。多个用户向量组成了用户－行为特征权重矩阵,其中,列为用户向量,行是行为特征,两者交叉的值是某特征在某用户中的权重,权重值根据 TF－IDF 算法所得。

将该矩阵行列转置后存储到 PostgreSQL 中,作为用户相似度计算和聚类分析的输入数据,如图 3 所示:

user_id integer	t1197 double preci	t1218 double preci	t1219 double preci	t1226 double preci	t1227 double preci	t1228 double preci	t1229 double preci	t1230 double preci	t1231 double preci	t1232 double preci	t1237 double prec
491	0.005360906	0.038803115	0.003345804	0	0	0	0	0	0	0	0.007414407
398	0.000757712	0	0	0	0	0	0	0	0.012159828	0	0
493	0.016029637	0.025275959	0	0.033817057	0	0.034101088	0.018615147	0	0	0	0
399	0.002043063	0	0.002546309	0.011174505	0	0.01126836	0	0	0	0	0.002507867
400	0.006133079	0	0.009236227	0.03234675	0	0.032618433	0	0	0.013123215	0	0.001411570
492	0.03743061	0.044713322	0	0	0	0	0.009879092	0	0	0	0
401	0.000227539	0	0.000567174	0.028801858	0	0.029043766	0	0	0	0	0.000628437

图 3　用户－行为特征权重矩阵片段

3.4　聚类分析,识别人物角色

(1) 相似度计算

用户间的相似程度实际上是通过用户向量间的距离来度量的,向量间的距离越近则表明用户间的相似度越大,采用欧式距离[10]来计算,公式如下:

$$dist(\vec{x},\vec{y}) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \qquad (4)$$

其中,x,y 为 n 维空间中的两个点,dist( ) 为两点间的距离,其值等于两点在所有坐标维度上的差值平方的和的平方根。

(2) 聚类分析过程

①采用 K－means 算法,其初始参数的配置为:聚类类别数量 k 值,依次测试 k＝9－15 的情况,结合实际的用户访谈和高频特征分布特点分析,最终选定 k＝10;终止迭代的条件,最大迭代次数为 10(实际迭代次数要小于该值),收敛性标准为 0.02,迭代过程中满足其中一个条件,迭代就会终止。

②系统自动选择 10 个初始聚类中心。

③经过 5 次迭代,由于聚类中心的变化小于收敛标准 0.02,于是迭代终止、开始收敛。

④迭代终止后,每个群集的聚类中心也趋于稳定,形成最终聚类中心,如表 2 所示:

表 2　最终聚类中心片段

特征 ID	群集 1	2	3	4	5	6	7	8	9	10
1290	.00	.17	.00	.00	.00	.00	.00	.00	.00	.00
1180	.08	.00	.03	.01	.01	.00	.14	.01	.00	.00
1222	.00	.00	.01	.09	.02	.12	.00	.02	.00	.01
1182	.00	.00	.02	.15	.07	.32	.00	.19	.12	.02

⑤最终聚类生成 10 个群集。当用户向量与某一群集中心的距离最小时,该用户就被分入该群集,每个群集所包含的用户样本略。

4　实验结果分析

4.1　各群集之间的相近程度

最终聚类中心间的距离可反映各群集之间的相近程度,如表 3 所示:

表 3　最终聚类中心间的距离

群集	1	2	3	4	5	6	7	8	9	10
1		1.529	.805	.893	.832	1.411	1.208	.904	1.653	1.577
2	1.529		1.299	1.348	1.309	1.735	1.592	1.355	1.935	1.871
3	.805	1.299		.360	.173	1.149	.923	.374	1.442	1.329
4	.893	1.348	.360		.353	1.027	.998	.439	1.475	1.388
5	.832	1.309	.173	.353		1.144	.944	.312	1.447	1.360
6	1.411	1.735	1.149	1.027	1.144		1.480	1.155	1.824	1.763
7	1.208	1.592	.923	.998	.944	1.480		1.008	1.712	1.638
8	.904	1.355	.374	.439	.312	1.155	1.008		1.478	1.397
9	1.653	1.935	1.442	1.475	1.447	1.824	1.712	1.478		1.972
10	1.577	1.871	1.329	1.388	1.360	1.763	1.638	1.397	1.972	

由表 3 可知,群集 1、2、6、7、9、10 与其他群集之间的距离相对较远,距离值均大于 0.8,说明这几个群集的用户在行为模式上明显有别于其他群集的用户;群集 3、5 之间的距离最近,值为 0.173,说明这两个群集的用户在行为模式上较为相近,但同时又存在一定的差别;群集 4、8 与群集 3、5 之间的距离也比较近,距离值在 0.35 附近,这类群集的用户行为模式介于前两种情况之间。

4.2　各群集显著行为特征类型分析

各群集的行为模式由最终聚类中心的行为特征及其权重组成,数学表示见表 2。其中,特征代表行为模式中行为的内容,特征权重代表行为模式中各行为的显著程度。

那些权重很高的显著行为特征,在很大程度上决定了群集之间行为模式的差别。将表 2 中行为特征分别按其所在群集的权重由高至低排列后发现,权重比较高的某些显著特征常常同时出现在一个或几个群集中,它们之间存在一定的相关性,这类特征可以共同帮助用户达成某一项任务。据此,对显著行为特征类型进行划分,结果如表 4 所示。

表 4　显著行为特征类型划分

	序号	显著行为特征	特征类型
显著行为特征类型	1	browse – title	题名导航
	2	subject = 略, items – by – subject, order = title, subject – search	主题导航
	3	community – list; 1,14, 18, 9, 6, 2, 7, 4, 5, 15, 17, 33, 35	社群导航
	4	browse – author, items – by – author	作者导航
	5	browse – date, order = DESC, sort_by = 2	日期导航
	6	simple – search, query = , query = 略; order = DESC, start = 10, sort_by = 2, rpp = 10, sort_by = 0	检索功能
	7	bitstream	文献下载
	8	2942, 1170, 略. pdf	首页推荐的文献资源
	9	mydspace, password – login, logout	个人空间
	10	mycustomize	个性化定制
	11	quick – submit, submit_authors = 略, tools, edit – item, checkTitle, submit = 略, subject = false, dc_title = , content – stat	提交个人文献
	12	POST, HEAD	Web 资源请求方法
	13	dspace – admin(系统管理员); authorize(授权); content – stat(统计); tools, edit – item, edit – epeople, eperson – list, group – edit, edit, edit – communities, mode = full, submit_simple = 略, submit = 略, multiple = false; (编辑)	系统管理
	14	admin_login. asp, admin, manage, manager;(系统登录) webeditor,editor,eWebEditor,edit,upfile_flash. asp,htmledit,eWeb,ewebeditor,WebEdit, include, htmleditor(系统开发)	系统开发

表 4 得出 14 种显著行为特征类型,涉及 IR 的导航、检索、下载、个性化、提交个人文献等系统功能,被高度关注的文献资源以及特别授权的系统管理、系统开发相关功能。

4.3　各群集行为模式的文字描述及用户数量分布情况

为方便读者直观理解各群集的行为模式,在其数学表示基础上,结合显著行为特征及其类型的划分,采用文字描述的方式进一步解读。此外,通过各群集所包含的用户数量、占用户总量的百分比两个指标区分主、次角色,以判断 IR 中各群集的重要性。

各群集行为模式与一种或几种类型的显著行为特征按其重要程度存在一定的对应关系。比如对群集 4 来说,权重高的前 13 个显著行为特征,分属于第 9 类个人空间和第 13 类系统管理中的编辑类特征(见表 4),占据该群集行为模式的主导地位;而权重相对较低的 5 个特征,分属于第 1 类题名导航、第 4 类作者导航、第 7 类文献下载和第 11 类提交个人文献特征,占该群集行为模式的非主导地位;据此,将该群集的角色定位为"编辑型的系统管理员",其行为模式是以系统管理员的身份登录个人空间,侧重编辑类系统管理功能的使用,对题名、作者导航等功能进行一般性使用。其他群集采用同样的分析方式,分析结果的文字描述,如表 5 所示:

表 5 各群集行为模式的文字描述及用户数量分布情况

		用户数	百分比	显著行为特征类型	行为模式的文字描述
聚类群集	1	2	.9	8,6,3	了解系统功能型:下载首页推荐文献,使用检索功能,几乎点击每个社群导航
	2	6	2.8	14	系统开发型
	3	86	40.4	6,2,1,11,10	关注主题内容的综合型用户:偏爱简单检索、主题导航和题名导航一类侧重主题内容发现的功能,对提交个人文献、个性化定制等功能进行一般性使用
	4	15	7.0	9,13,1,4	编辑型的系统管理员:以系统管理员的身份登录个人空间,侧重编辑类系统管理功能的使用,对题名、作者导航等进行一般性使用
	5	56	26.3	4,11,3	关注作者与提交个人文献型:侧重对作者导航和提交个人文献功能的使用,对社群导航进行一般性使用
	6	2	.9	9,13	授权型的系统管理员:以系统管理员的身份登录个人空间,侧重授权系统管理功能的使用,对编辑类管理功能进行一般性使用
	7	1	.5	6,3	侧重检索和按某社群导航浏览型
	8	33	15.5	9,11,6,3,4	单纯的提交个人文献型:以普通用户的身份登录个人空间,侧重个人文献提交功能的使用,偶尔使用简单检索和按社群、作者导航浏览功能
	9	11	5.2	14	登录型的系统开发人员
	10	1	.5	2	单纯的关注主题内容型
	共计	213	100		

由表5可知,群集3、5、8所占百分比最高,三者累积百分比达82.2%,为系统服务的主要角色,他们表现出对主题内容、作者和提交个人文献等相关功能频繁使用的行为模式;群集2、4、9所占百分比次之,三者累积百分比达15%,为系统服务的次要角色,他们的行为模式属系统管理和系统开发类;群集1、6、7、10的用户数量很少,1-2人左右,他们在个别行为特征上非常显著,为系统服务的个别角色。

5 人物角色量化模型的数学表示

将各群集行为模式的矩阵实例(即最终聚类中心,见表2)抽象化后,得到人物角色量化模型的矩阵表示如下:

$$A_{m\times n}=\begin{array}{c} \\ t_1 \\ t_2 \\ \cdots \\ t_m \end{array}\begin{array}{c} \begin{array}{cccc} u_1 & u_2 & \cdots & u_n \end{array} \\ \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \end{array} \qquad (5)$$

其中,$A_{m\times n}$代表人物角色-行为特征矩阵,n=10,m=213;每列为人物角色的向量,每行为行为特征的向量;a_{mn}为第n个行为特征在第m个人物角色中所占的权重。

6 结 语

人物角色量化模型是基于用户行为模式而对用户群体类别做出的划分,不论从数据来源和分析方法都具备了客观性,而经过量化后的模型以矩阵形式表示使得该模型可以以数字形态进行存储和使用,使得数字化服务系统在提供个性化服务时可以方便地调用该模型,进而支持动态的个性化服务。而通过程序实现该模型的生成,使得人物角色模型可以方便的生成和更新,同时该方法和模型具备一定的通用性,可以适用于大多数的服务系统。

本文在机构仓储系统的应用环境下,通过对用户行为日志的分析,采用聚类方法创建人物角色量化模型,识别出具有共同行为模式的用户群体类型,并以矩阵形式来表现该模型。但研究中仍有不足之处和需要继续深入研究的方面,如用户识别的规则比较简单,还有相当数量的用户没能区分开;聚类类别数量根据人为经验确定,可结合因子分析等统计方法计算得出;行为特征颗粒度的选择还比较粗;应在聚类过程中加入对相同特征变量数量的限定等。

参考文献:

[1] OCLC Research - JISC Report[EB/OL]. [2010-06-17]. http://www.oclc.org/us/en/news/releases/2010/201026.htm.

[2] Cooper A. 交互设计之路——让高科技产品回归人性[M]. Ding C 译.2版. 北京:电子工业出版社, 2006.

[3] 孙敏杰,吴振新. 人物角色方法及其在数字图书馆的应用初探[J]. *图书馆杂志*,待发.

[4] 祝忠明,马建霞,张智雄,等. 中国科学院联合机构仓储系统的开发与建设[J]. *图书情报工作*, 2008,52(9):90-93,144.

[5] Web Characterization Activity[EB/OL]. [2010-06-17]. http://www.w3c.org/wca.

[6] List of User-Agents (Spiders, Robots, Crawler, Browser)[EB/OL]. [2010-10-09]. http://www.psychedelix.com/agents/index.shtml.

[7] TF-IDF 百度百科[EB/OL]. [2010-10-09]. http://baike.baidu.com/view/1228847.htm#3.

[8] Sinka M P, Come D W. A Large Benchmark Dataset for Web Document Clustering[A]. *//Soft Computing Systems: Design*, Management and Applications, Frontiers in Artificial Intelligence and Applications[M]. 2002:881-890.

[9] 庞景安. 科学计量研究方法论[M]. 北京:科学技术文献出版社, 2002.

[10] Euclidean Distance[EB/OL]. [2010-10-09]. http://en.wikipedia.org/wiki/Euclidean_distance.

(作者E-mail:sunminjie@mail.las.ac.cn)