分众分类系统中的用户行为研究*

赵文娟[1]，张宁[2]，吴真明[3]
（1. 山西大学商务学院，太原 030031；2. 山西大学管理学院，太原 030006；
3. 北京四方继保自动化股份有限公司，北京 100085）

摘要：分众分类法Folksonomy作为一种平民分类法是Web 2.0时代的重要产物，为网络知识的组织和共享提供了一个良好的途径。分众分类系统中，用户通过标签Tag对网络资源进行标注，标签特征一定程度上反映了用户的行为特征，因此文章首先通过对用户标签的统计分析，总结用户的标签特征，然后分别从标注动机、用户认知和社会认同角度对用户行为进行研究，最后根据用户的网络行为建立用户兴趣模型。

关键词：分众分类法；标签；用户行为；兴趣模型
分类号：G254
DOI：10.3772/j.issn.1673—2286.2014.08.004

1 引言

当前网络资源的组织和管理具有用户高度参与性的特征。2004年陆续出现的许多社会性软件，如分享书签网“del.icio.us”，相片分享网“Flickr”，目标分享网“43_Things”等，允许用户以任意关键字为标签，对其发布或使用的网络信息进行标注，以实现所标注资源共享，该类型系统又被称为分众分类系统，用户通过添加元数据，并使用标签组织他们获得的资源，为其分类。作为一种自下而上的分类系统，用户有能力积累自己的在线经验，并使得其他用户快速检索到所分享的资源，以获得更好的用户体验。

与LinkedIn、Friendster等网络站点不同，采用分众分类方法的网络站点更致力于组织数据、发展关系。用户在公开的网络环境中组织自己及他人的数据，使得用户可以共享数据并找到兴趣相投的人。用户建立标签的目的是能够比较快地找到他们所需要的信息，这是一种面向大众的记忆方法，该方法只与记忆有关，无关分类是否正确，也无关分类的准确性和权利，是按照个性需求而分类，同样服务于网络上所有用户。在整个过程中，充分体现了用户的主动性和个性化，而且能参考和借鉴其他用户所选择的资源和标签，体现了交互性和社会性。对用户行为的研究，对进一步揭示分众分类系统标注动机、方式和质量有很大的意义[1]。

文章首先通过对用户标签进行统计，总结标签所呈现出来的规律和特点，揭示通过标签所折射的用户行为特征，然后通过对用户的统计及问卷的汇总分析，从用户的角度分析用户在标注动机、认知及社会认同方面的行为特征，最后根据用户的网络行为构建用户兴趣模型。

2 用户标签特征

分众分类系统中，标签连接信息资源、信息发布者和信息用户，形成关系网络的关键节点[2]。用户对网络资源进行标注、分类和索引，通过标签揭示和挖掘内容，是实现大众标注的核心要素和手段，因此对标签规律的研究能在一定程度上反映用户的行为特征。

2.1 数据来源

对《中国分类主题词表》第五版教育类目下对应的主题词进行筛选、去重、频率统计等处理后得到1059个主题词，人工剔除低频率词（出现频次少于等于2）、字面意思与教育关联不大的词，得到的322个主题词。以这322个主题词为检索入口，经过去重，英文大小写合

* 本研究得到国家社会科学基金青年项目“基于框架网络本体的标签系统语义分析研究”（编号：13CTQ030）资助。

并后在“del.icio.us”网站获得11415个不重复的标签数据集。在获得的11415个标签中，包括有日、韩、英、中四种语言标签及少量特殊符号标签。本文只对中文标签进行分析研究。在获得的所有中文标签中包括5677个纯中文标签和57个噪音中文标签，噪音中文标签指中文加英文，如高校BBS；中文加数字，如医学2.0；中文加特殊符号，如社会.政治.经济。

2.2 数据统计

标签的统计包括标签的频次分布、标签类型分布及标签特征。

（1）标签频次分布

首先对标签对应的资源数量进行统计，发现“教育”标签使用频率最高，资源数量4117个，其次是“教材”，资源数为2421个，“大学”位列第三，资源数量为1652个，但是只涉及一个资源的中文标签的数量最多，约3464个，几乎占到了全部标签数的61%，统计表见表1。

表1 标签-资源数量的分段统计表

分类标准	标签所涉及资源数量	标签数	百分比
独立标签	1	3464	61.02%
简单分享标签	2	766	13.49%
简单网络标签	3	378	6.66%
复杂网络标签	x≥4	1069	18.83%
合计		5677	100.00%

（表中，独立标签：只有标记一个资源的标签；简单分享标签：标记两个资源的标签；简单网络标签：标记三个资源的标签；复杂网络标签[3]：标记四个以上（包含四个）资源的标签，它属于使用频率较高的标签。）

统计结果表明，用户对于标签的选择一定程度上受到“从众心理”的影响，往往选择高频词标签去标注网页，比起来自叙词表的专业术语，这些大众化的词语更容易被用户接受，并通过用户自己的标注影响其他用户标签的选择。越来越流行的高频词构成了核心标签集，低频词则由于资源和用户的逐渐稀少而被边缘化[4]，说明标签频次遵循了“马太效应”。所占比重很小的高频标签标注的资源数量却在整个资源数量中占据了很大的比例，遵循了经济学中的“二八定律”。

（2）标签类型分布

标签类型主要指标签长度、简称和特殊标签三种类型。

标签长度指的是标签所含文字的个数。对收集到的资源，按照标签长度不同进行统计，得到表2。

表2 标签长度统计表

标签长度	1	2	3	4	≥5	合计
资源数量	26	3339	1132	875	305	5677
百分比	0.46%	58.82%	19.94%	15.41%	5.37%	100.00%

结果显示长度为2-3的标签占了绝大多数，这些标签多为长度为2或3的单一词，说明在分众分类系统中，用户习惯于用单一词语标签来标注和组织信息，从某种程度上说，分众分类系统中的关系网络聚类功能是通过使用单一术语标签来实现的。

简称是由长的、复杂的词语压缩简化而成的短的、简单的词语。中文标签中的简称集中在二字简称和三字简称，简称标签共80个。统计结果如表3。

表3 简称统计表

标签	标签数	资源数
二字简称	65	492
三字简称	15	78

简称标签的使用说明用户在组织、获取和利用资源时遵循“最少付出原则”或“最小努力原则”，用户期待用最小的努力最大限度地表达自己的知识或观点。

特殊标签也称复合词标签，指用空格符将一个标签分成几个部分，其实质是用户用独立的几个词的组合作为一个完整的标签来标注资源。这样的标签在整个中文标签当中共有18个。特殊标签的使用显示出分众分类系统中用户的友好性，用户自由表达自己的观点、想法和对资源的情感或判断，为资源的分享提供了一个良好的平台。

标签类型的统计结果表明，多数用户没有受过专门的培训，并不是信息组织和检索方面的专家，因此标签的选择偶尔也会使用多字单词标签和常规简称。

（3）标签特征分析

在中文标签中，除了标签词性特征，经常还会考虑标签是否可切分。在收集到的5677个中文标签中，有

58.78%的标签为不可切分标签，数量约为3337个，剩余的2340个标签为可切分标签，占标签总数的41.22%。在不可切分的标签集合中，名词所占比重最大，为66.25%，其次为动词，所占比重为27.66%，形容词位居第三，所占比重为4.43%，其余词性标签所占比重为1.66%。加拿大学者Margaret E. I. Kipp[5]认为，分众分类系统中用户使用的标签分为情感性标签和时间任务相关性标签。情感性标签多为形容词，时间任务相关性标签多为名词或名词短语。统计结果显示，时间任务型标签占主导，约为66%，情感性标签所占比重很小，只有4.43%。

2.3 用户标签的特点

通过对用户标签统计结果的分析，得出用户标签有如下特征：

（1）标签体现大众智慧

区别于传统的分类法，分众分类中的标签不具有受控性和层次性，标签体现的是个人词表，反映了个人的需求，不受不同文化、政治和社会形态的影响。分众分类法汇集了每个个体的观点，标签对应资源的频次分布遵循“长尾效应”，多数标签位于尾部，特定标签远远超过高频标签。

（2）标签的及时性

基于标签的分众分类系统提供了一个灵活及时的分类体系。当用户标注感兴趣的网页资源时，可立即创建标签，并对资源进行收藏和共享。这种灵活性可以使用户对网络资源语言选择的偏好和网络信息的动态变化迅速地做出回应，使分众分类法在传统的受控分类法面前具备了很强的竞争力。一个受控词表的建立，可能要花费几年的时间，并且形成的词表结构固定，层次分明。而面对不断变化的事物，人们的表达方式也随之变化，标签使分类系统容纳了充分的新词汇、新术语。

（3）标签体现了用户需求和愿望

标签体现的是用户直接的信息需求和愿望，反映的是用户对信息的分类方式。例如一张猫的图片，可以有标签如猫、小猫、虎斑猫、可爱、机灵，不同的用户根据自身对图片的期待选择标签，在传统分类法中，将这多个标签概念合并成一个术语，会发生信息丢失，并忽略了用户情感。分众分类系统在用户浏览相关标签主题和其他用户资源时，同样促进了用户进一步地学习和探索，多数用户在浏览或检索资源时，并不能十分确切地表达自己想找什么样的资源，标签使得用户不通过刻意的方式就能够找到其他途径，发现新的资源。

3 用户行为研究

相比标签，直接对用户进行研究更能直接体现用户行为特征。文章从标注动机、用户认知和社会认同[6]三个方面对用户行为进行研究。

3.1 数据获得

对用户行为研究的数据的获得通过两种途径：一、问卷；二、2.1中标签的用户统计。问卷的目的主要是统计用户的标注动机、影响标注动机的因素、获取资源的途径、对标注价值的认知。为方便问卷者，全部为选择题。问卷主体分为两部分，第一部分基本信息包括年龄段、学历、有无网络标注经历，单选。第二部分为不定项选择，包括标注动机是什么、对不同的资源的标注动机等，共15题。问卷采用QQ邮箱好友推送的方式，收回有效问卷261份。对2.1中标签用户的统计主要为了分析用户的标注态度及社会认同与标注行为间关系。

3.2 标注动机

标注行为是指利用标签对网络信息资源进行标注，以供标注者自身或他人检索信息资源的过程或结果。而动机是人为满足特定的需求所产生行为倾向的原因。

标注动机分为无动机、外在动机和内在动机，无动机指用户进行标注时没有自己的主观意识。外在动机包括互动与竞争、分享、唤起关注，内在动机包括表达情感、阐述观点、组织资源、记忆学习[7]，其中内在动机是用户标注行为的决定性因素。对收回的261份问卷进行统计，34份为从未实施过标注活动，约占统计人数的13%，57份为无动机，约占统计人数22%，其余170份都有较为明确的标注动机，详情如表4。

由表4可见，多数人标注的目的为组织资源、记忆学习和分享。在实际的标注行为中，用户的标注动机往往不是单一的，而是可能包含有多个动机，因此会有总体比例之和大于1的情况。此外，标注行为中存在一些因素影响着用户的标注动机，详见表5。

表4 用户标注动机统计表

动机	无动机	表达情感	阐述观点	组织资源	记忆学习	互动竞争	分享	唤起关注
份数	57	18	106	198	185	74	135	39
比例	25%	8%	47%	87%	81%	33%	59%	17%

表5 影响标注动机的因素

影响因素	例子
站点类型	专业网站：组织资源、分享；购物网站：阐述观点、分享；社交网站：分享；学术网站：记忆学习等
资源类型	图书：组织资源、记忆学习；博客：分享、唤起关注；商品：表达情感、阐述观点等
用户信息素养	较高：唤起关注、表达情感、记忆学习；一般：组织资源、分享、记忆学习；较低：阐述观点、表达情感
行为强度	分享为动机的行为强度最大，唤起关注为动机的行为强度最弱
行为频率	组织资源、阐述观点为动机标注最频繁，唤起关注为动机的标注频率最低
标签便利性	很方便：阐述观点、记忆学习、组织资源；比较方便：记忆学习、分享；不方便：记忆学习、阐述观点
标签类型	情感性标签：表达情感、唤起关注；时间型标签：记忆学习、组织资源；任务型标签：阐述观点、组织资源
标签形式	长标签：组织资源、记忆学习；简称标签：记忆学习；符号标签：表达情感、唤起关注

3.3 用户认知

用户认知主要包括用户获取利用资源情况、用户的标注态度及行为特征、用户对标注作用和价值的认知面[8]。

3.3.1 用户获取资源情况

用户获取利用资源情况指用户更趋向于以何种方式获取资源。以教育领域为例，我们获得如下的结果，见表6。

用户信息获取的途径与用户年龄有很大的关系，年龄小于20的用户，主要为中学生，获取知识的主要途径比较单一，主要是课堂学习。年龄在20到40岁的用户，构成网络标注用户的主体，这一类人群学习目标明确、精力旺盛、有很强的组织学习及与人分享的欲望，他们是网络上的活跃分子，获取信息的主要途径为网络阅读和网络下载。40岁以上的用户多为不惑之年的中年人，他们中很少一部分为某一领域的专家，通过网络阅读随时跟进领域前沿。但是绝大多数用户只是通过网络阅读打发寂寞时光，他们阅读的主题很泛，阅读行为随意。

表6 用户年龄与获取资源途径表

用户年龄	主要途径
Age<20	课堂学习、网络阅读
20≤Age<30	图书馆、网络下载、网络阅读
30≤Age<40	网络阅读、网络下载
Age≥40	网络阅读

3.3.2 用户的标注态度及行为特征

用户的标注态度指用户有多少意愿愿意主动参与标注，用户行为特征指标注行为的标签更趋向于表达何种动机。

用户统计发现，近一半的用户只是偶尔标注，这部分用户构成普通用户群，约三成用户为很愿意标注，他们是网络标注中的活跃分子，构成活跃用户群。其余两成用户为从不标注，被称为“看客”。用户的受教育程度也在很大程度上影响了用户信息获取的方式和标注行为。同一年龄段当中的用户，高学历者更愿意参与到网络资源的组织和他人共享的社区中，低学历者更趋向于只“看”不“标”。

在用户的标注行为中，标注动机排名情况为：分享、更好发现网络资源、揭示资源的主题内容、更好地

组织资源、促进协同学习、评价等。

3.3.3 用户对标注作用和价值的认知

标注作用指标注行为是否能有效地促进资源的组织、发现和共享。近四成用户认为标注能有效地促进学习，约四成用户认为能在一定程度上促进学习，还有不到一成的用户认为标注没有任何价值。

标注对学习的促进作用主要体现如下几个方面：能够聚类同一主题资源；能够扩大检索范围，更有利于发现新资源；能够促进资源共享。

3.4 社会认同度

社会认同理论是人们进行是非判断的标准之一。就是在做出决定时考虑别人是怎么想的，这一理论尤其适用于在不确定的环境中行为人的决策，面对未知的事件，很自然地环顾周围了解其他的反应[9]。

我们根据用户的被关注人数将用户进行分类。将被关注用户数超过10的用户定义为活跃社交型用户，被关注用户数在1-9的用户定义为普通社交型用户，不被任何用户关注的用户定义为普通用户。在收集到的587名用户中，三类用户分布情况详见表7。

表7 用户被关注人数统计表

用户类型	被用户关注的人数	用户数	百分比
活跃社交型用户	≥10	63	10.73%
普通社交型用户	1-9	171	29.13%
普通用户	0	353	60.14%
总计		587	100.00%

从表中可以看出，约60%的用户并没有与其他用户建立联系，他们只关注网络资源，还没有关注到其他用户，构成用户的主体。不到11%的用户为活跃的社交型网络用户，他们熟悉并经常使用和分享网络资源，并主动关注他人，这部分用户的标签和资源都有很高的质量，是某一领域的热爱者或者权威人士。

4 基于用户行为的兴趣度模型的构建

用户兴趣模型的建立是通过发现、记录并分析用户网络行为，准确描述用户兴趣，构建用户兴趣模型，将用户真正感兴趣的资源推荐给用户，同时在用户之间进行准确的推荐[10]。

分众分类系统中的标注都由三个实体构成：标签、用户和资源。兴趣度模型通过建立用户集、资源集，将用户和资源建立映射，通过用户浏览标注的资源得到用户感兴趣的资源类别，结合考虑用户行为及权重构建。

4.1 网络资源挖掘

网络资源分为服务器资源和客户端资源。当用户在某一站点访问一次，网络服务器中会增加一条访问记录，包括用户IP、用户名访问时间、访问资源的URL等，在客户端服务器中则获取如下信息：用户的浏览路径、用户的浏览行为、在某一资源上的浏览时间等。

对服务器端资源的挖掘可得哪些用户正在访问哪些资源、用户的访问路径、被访问的资源名称、访问页面的时间。对客户端资源的挖掘可得当前用户正在访问的资源、用户访问资源的集合、用户的浏览时间、用户访问URL的累积时间、用户的访问行为等。

4.2 用户行为分析

用户对资源的浏览行为体现在三个方面：标记行为、操作行为和重复行为。我们将其定义为集合S，S={A1，A2，A3}，A1代表标记行为：添加tag、删除tag、保存资源、打印资源。A2代表操作行为：复制、粘贴、剪切、点击链接等。A3代表重复行为：指对统一资源的重复访问。

用户的兴趣行为包括平均浏览时间、访问的同一tag、重复访问同一资源、点击链接、保存标签。我们将用户的兴趣行为定义为集合L，L={B1，B2，B3，B4，B5}，其中B1：平均浏览时间，B2：访问的同一tag，B3：重复访问同一资源，B4：点击链接，B5：保存标签。在模型中，将5个元素分别作量化处理，并根据对用户兴趣度的影响赋予不同的权值，所有元素权重总和为1。

用户的浏览行为定义为集合T，T={C_1，C_2，C_3…..C_n}，其中C_i=<t_i,c_i,B_i>，代表用户一次成功的浏览行为，t_i表示B_i发生的时刻并按照时间序列排序，c_i表示用户的兴趣内容，B_i表示用户兴趣行为。

4.3 用户兴趣模型的构建

在用户的兴趣模型中，用户的兴趣向量为网络中的节点，每一个兴趣节点的权值代表兴趣度的大小，通过用户的浏览行为和浏览内容共同得到用户的兴趣模型，流程如图1。

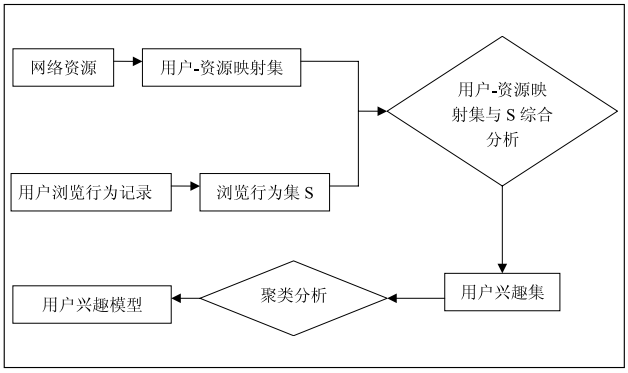

图1 用户兴趣模型构建流程图

5 结语

分众分类法为知识的组织和共享提供了一个良好的途径。本文通过对标签、用户、资源的统计，揭示了分众分类系统中用户标签的特点、用户标注行为特征及规律，并参考用户的网络行为提出了用户兴趣模型的构建流程。

参考文献

[1] 杨青云，裴雷，吴克文. 国外社会化标注系统中标注行为研究现状[J]. 情报杂志，2009(11)：185-188

[2] 查先进，吕彬. 知识共享视角下的大众标注行为研究：基于标签的实证分析[J]. 图书馆论坛，2010(6)：76-80

[3] 张宁. 分众分类系统的用户行为特征分析[D]. 太原：山西大学，2013

[4] VANDER WAL. Optimizing Tagging UI for People[J]. People & Search，2010：165-178

[5] NORUIZ A. Folksonomies:controlled vocabulary[J]. Knowledge Acquisition，2007(18)：37-45

[6] HOTHO A，JASCHKE R，SCHMITZ C. A Ranking Algorithm for Folksonomies[J]. Gerd Stumme FolkRank，2010(8)：76-83

[7] 王娜，马云飞. 网络环境下大众标注行为动机的调查和分析[J]. 图书情报工作，2013（12）：100-106

[8] 魏来，王雪莲. 社会标注在学习资源组织中的应用及用户认知调查[J]. 情报杂志，2013（5）：185-189

[9] BINKOWSKI P J. The Effect of Social Proof on Tag Selection in Social Bookmarking[J]. Applications，2009(5)：190-201

[10] 王微微，夏秀峰，李晓明. 一种基于用户行为的兴趣度模型[J]. 计算机工程与应用，2012,48（8）：148-151

作者简介

赵文娟，女，1983年生，硕士，山西大学商务学院讲师，研究方向：计算机信息检索、语义标注，E-mail：zhaowenjuan1118@163.com。
张宁，男，1986年生，硕士，山西大学经济与管理学院，研究方向：知识组织。
吴真明，男，1981年生，工程师，北京四方继保自动化股份有限公司，信息系统应用部咨询顾问，研究方向：企业资源管理。

Research for the Users’ Behavior in the Folksonomy System

ZHAO WenJuan[1], ZHANG Ning[2], WU ZhenMing[3]
(1. Business College of ShanXi University, Taiyuan 030031,China; 2. School of Economics and Management, Shanxi University, Taiyuan 030036, China
3. Beijing Sifang Automation Co,LTD. Beijing 100085,China)

Abstract: As a civilian classification, Folksonomy is an important production of the era of the Web2.0, it provides a good approach for the organization and sharing of network knowledge. In the Folksonomy system, users through the tags to labeling web sources. Through the analysis of user tags, looking for the statistics distribution, summarize the users tags action, study the users behavior respectively form tagging motivation, cognitive and social identity, and finally build users interest model based on users behavior.

Keywords: Folksonomy; Tags; Users’ behavior; Interest model

（收稿日期：2014-07-04）