用户社会化标注中非理性行为的表现及原因分析

林鑫[1]，梁宇[2]
（1.华中师范大学信息管理学院，武汉 430079；2.国网鄂州供电公司电力调度控制中心，鄂州 430060）

摘要：为深化对用户社会化标注行为机制的认识，推动标签应用研究与实践发展，本文采用日志分析法和深度访谈法对用户社会化标注非理性行为的表现和原因进行研究。结果表明，非理性社会化标注行为是一种普遍存在的现象，主要表现在对标注对象选择、标注角度选择和标签表达三方面；其原因主要与用户态度、标注动机、标注习惯及社会化标注系统四个方面有关。

关键词：社会化标注；标签；用户；非理性行为
中图分类号：G254
DOI：10.3772/j.issn.1673-2286.2016.12.008

作为Web 2.0的一项典型应用，社会化标签受到广泛关注，并被应用于检索[1-2]、资源分类和聚类[3-4]、个性化推荐[5-7]等领域。这些研究与实践的开展隐含两个前提假设：一是资源标签能较全面、准确地反映资源内容和形式特征，二是用户标签能较充分、准确地体现其兴趣偏好。然而，这两个假设并未得到证实。相反，用户在社会化标注中经常存在一些非理性行为，即在进行标注时非常随意，标注行为的产生缺乏依据，典型表现是相似情境下标注行为区别显著。以社会化标签系统“豆瓣电影”为例，用户“春衫薄透”“John、”和“houge1731”（数据来源：豆瓣电影网）均看过类型、主题、演员、风格非常类似的《谍影重重》系列影片。而在标注中，用户“春衫薄透”仅对《谍影重重3》添加标签；“John、”虽对3部影片都添加了标签，但仅对《谍影重重1》添加标签“间谍”，《谍影重重2》只添加标签“动作”，为《谍影重重3》仅添加标签“特工”；“houge1731”在标注男主角名字时，采用“马特·达蒙”和“Matt_Damon”两种拼写方法。因此，针对用户社会化标注中的非理性行为进行研究，分析其表现和原因，既有利于深化对用户社会化标注行为机制的认识，也有助于推动标签应用研究与实践的发展。

1 相关研究现状

自标签产生之初，社会化标注中的非理性行为就受到关注，但已有研究主要集中于词汇选用和拼写方面；同时，部分研究围绕影响标注或标签选用的因素展开，其研究成果也有助于解释社会化标注中的非理性行为。下面将从两方面对相关研究进行综述。

1.1 标签词汇选用和拼写非理性行为

关于标签词汇选用和拼写非理性行为的研究，在初期侧重于非理性行为的证实和原因分析，目前则更侧重于非理性行为引发问题的解决。

标签词汇选用和拼写非理性行为的证实和原因分析。标签词汇选用和拼写非理性行为主要表现在歧义词多、同义词多、单复数混用、缩写、主观性标签拼写错误和私人标签方面[8-11]。这也是社会化标注中较为常见的问题，Kipp[12]、Thomas[13]、查先进等[14]分别以不同的社会化标签系统进行证实。其原因主要包括标注系统缺乏词汇控制，允许用户自由标注；标注系统缺乏标注规范指引；用户标注态度随意[3]。

关于标签词汇选用和拼写非理性行为引发问题的应对策略，部分学者建议规范标签使用，具体措施包括链接专业主题词表供用户参考、允许用户反复编辑标签[12]，规范并指导用户标注图书、合并图书编目生成的元数据与图书标注生成的标签、建立个性化半自动标引[15]，在标签推荐时引入受控词表等[16]；还有学者建议在标签应用中采取策略加以解决，较受关注的思路是基于标签的本体构建以及标签与本体的关联[17-18]。

1.2 用户社会化标注影响因素

综合已有研究成果，影响社会化标注的因素主要包括动机、历史标签、系统推荐标签、社区内其他用户的影响等。典型研究包括Strohmaier等将标注动机分为资源描述和资源组织，并认为资源描述动机的用户更易使用私人标签，在标签使用行为上伴随更强的随意性；资源组织动机用户更倾向于使用规范的标签，且对各资源的描述角度比较一致[19]。Mirzaee等也认为动机与用户的标注行为有密切关系，并基于问卷调研的数据，构建标注动机与标注行为的关系模型[20]；Sen等认为主要影响因素包括用户历史标注行为、其他用户的影响、系统内置的标签选择算法[21]；统计发现，随着时间推移，同一个资源各标签的比例逐渐趋于稳定，Golder等由此认为社区内其他标注者是重要的影响因素[22]；Binkowski认为用户标注时会受到系统推荐算法的影响，尤其是对内容复杂、了解不深入的资源[23]。

综上所述，相关研究已注意到用户社会化标注中非理性行为，而关于标注行为影响因素的研究也为用户非理性行为的原因揭示提供依据，因此本研究的开展具备一定的基础。然而，当前关于非理性社会化标注行为的揭示仅限于词汇选用和拼写方面，且集中于用户群体视角，因而不够充分、系统。基于此，本研究将在延续群体视角的基础上，融入用户个体视角，并从标注对象选择、标注角度、标签表达三方面分析社会化标注中的非理性行为，并基于用户访谈探究其原因、基于用户日志数据量化分析其影响。

2 研究方法与数据

2.1 研究方法

本研究以知名社会化标注系统豆瓣电影的用户为样本，综合运用日志分析法和深度访谈法对用户社会化标注中非理性行为的表现和原因进行分析，研究流程简述如下。

（1）基于用户日志的非理性标注表现的假设提出。如果用户标注行为是理性的，则其每次标注行为都应有据可依，且在标注中遵循统一规则；因而，如果对比用户在不同时间段或针对相近影片的标注行为，一旦存在较多不一致情况，则可将其作为非理性标注的疑似行为。基于该思路，随机抽取20位用户的标注数据进行人工分析，结果显示，存在行为不一致的情况有三类：①标注影片选择存在不一致，典型表现是只标注同系列影片的一部分；②标注角度存在不一致，典型表现是有些影片标注制片国家/地区，有些则未标注；③标签选用和拼写存在不一致，典型表现是同义标签、错误拼写现象较多。此外，在标签分析中，也存在不少无法理解其含义的标签，或与影片不相关的标签。如对影人周星驰未参与的影片添加标签“周星驰”，未避免非理性行为表现遗漏，在初期也将该现象纳入非理性标注行为的假设。由此，基于日志分析形成标注对象选择、标注角度、标签选用和拼写、无意义或不相关标签的随意添加四个假设。

（2）基于用户非理性标注表现假设的深度访谈设计与实施。访谈目的是验证假设是否成立，对非理性标注表现的概括是否完整，以及在此基础上的原因探究。因此，访谈主要围绕以下十方面展开：①对社会化标注的认知和态度；②社会化标注的动机；③标注流程和标注时主要考虑因素；④只标注部分感兴趣影片的原因；⑤标注角度选择中非理性行为的原因；⑥添加的标签是否均认为与影片相关，并有明确含义；⑦选用不规范词汇进行标注的原因；⑧已注标签存在同义词或近义词的原因；⑨标签出现拼写错误的原因；⑩一些明显相关的标签未添加的原因。其中，对③—⑩结合用户感兴趣的多部影片进行访谈。

（3）基于用户访谈和日志数据梳理非理性标注行为的表现。基于访谈结果，对（1）中提出的假设进行验证，并探索是否存在新的非理性标注行为。在确认非理性标注行为表现的基础上，基于日志数据对非理性标注现象进行量化分析。

（4）基于用户访谈分析非理性标注行为的原因。结合社会化标注中非理性行为的具体表现对访谈数据进行开放式编码、关联编码和选择编码，最终将社会化标注中非理性行为的原因概括为用户态度、标注动机、标注习惯和社会化标注系统的影响四方面。

2.2 数据

研究所用数据包括用户日志和访谈，下面分别加以说明。

（1）用户日志数据。鉴于豆瓣电影不支持随机用户抽样，因此本文采集50位明星用户的关注列表[24]，获得830 682位用户ID，并借助工具采集每位用户的标注数据。在此基础上，从中随机抽取400位，作为本研究的样本，其中进行过社会化标注的用户为287位。这287位用户的数据就构成本研究的用户日志数据，采集字段包括用户ID、用户标记为“看过”影片的URL、影片名、影片基本信息、标记时间、标签内容。在研究实施中，用户日志数据被随机分成两组，一组由20位用户构成，进行人工分析提出非理性标注行为表现的假设；另一组由267位用户构成，进行用户非理性社会化标注行为的量化分析。

（2）用户访谈数据。这部分数据的用途是验证用户非理性标注行为表现的假设是否成立，并探索是否存在新的非理性标注行为现象，以及非理性标注行为产生的原因分析。访谈采用半结构化方法，抽样方面采用目的性抽样和滚雪球式抽样相结合的方法，在样本数量选择上依据信息饱和原则[25]，最终得到15位用户的访谈数据。这些受访者均为武汉大学学生，使用豆瓣电影时间超过1年，标注电影超过50部，属于社会化标注的熟练用户。其中，13位受访者存在非理性标注行为，2位不存在非理性标注行为。

3 用户社会化标注中非理性行为的表现

通过对15位用户访谈，证实了用户在标注对象选择、标注角度、标签选用和拼写三方面确实存在非理性现象，但并不会随意添加标签，即前文基于日志分析提出的用户可能会随意添加无意义或不相关标签的假设被证伪。此外，访谈中未发现新的非理性标注行为类型。因此，可将用户在社会化标注中非理性行为的表现概括为标注对象选择中的非理性、标注角度选择中的非理性、标注词汇选用和拼写中的非理性三方面。研究将基于抽样数据对上述现象进行量化描述，从而了解社会化标注中非理性行为的普遍程度。

3.1 标注对象选择中的非理性

标注对象选择中的非理性指用户仅会选择部分感兴趣的资源进行标注，而且在选择过程中存在较强的非理性行为，这种现象在社会化标注中非常普遍。为量化该现象，本研究将用户标记为“看过”的影片视为其感兴趣的全部影片（鉴于可能存在部分用户看过的资源未被标记为看过，因而实际情况会比分析出来的数据更为严重），统计用户进行社会化标注的影片占其“看过”影片的比例（见图1）。从结果看，50.9%的用户进行社会化标注的影片不超过其“看过”影片的20.0%，68.5%的用户不足60.0%，仅有19.9%的用户超过80.0%。

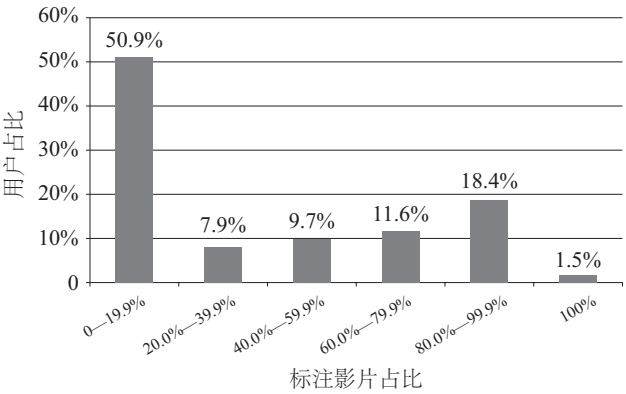

图 1 标注影片占用户“看过”影片的比例分布

3.2 标注角度选择中的非理性

如果用户的标注行为是理性的，则其应在标注中遵循统一的原则，具体到标注角度选择上也应该保持一致；若用户对不同影片的标注角度不同，则可认为其社会化标注是非理性的。为量化该现象的普遍性，以标注影片10部以上的用户为对象（若标注影片过少，则结果可能波动较大），统计其对制片国家/地区、类型/主题、年份三类标签的使用情况。选择这三类标签的原因有两个：第一，它们是用户常用的三类标签；第二，每部影片这三方面的属性均不为空。统计结果显示，在所有使用过制片国家/地区类标签的用户中，45.7%的用户标注制片国家/地区的影片的总数占其全部标注影片的比例低于40.0%，而1.7%的用户全部标注了制片国家/地区。其他两类标签的使用情况与其类似，具体数据如图2所示。

3.3 标注词汇选用和拼写中的非理性

用户标注时词汇选用和拼写中的非理性可区分为宏观和微观两个层面。宏观层面指从社会化标注系统

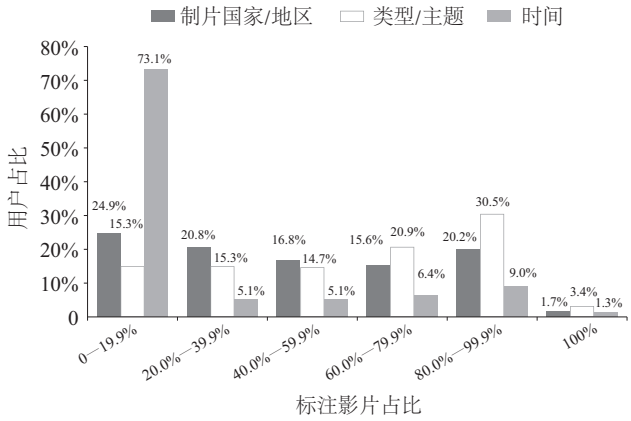

图 2 制片国家/地区、类型/主题和时间标签占用户标注影片的比例分布

的角度出发，所有不规范标签均可认为是标签选用和拼写中非理性的表现，包括私人标签、缩写、英文的单复数、同义词/近义词、复合词、错误拼写、简繁体等；微观层面指从单个用户角度出发，将个人标签集合内部出现的标签不一致和拼写错误问题视作其标签选用和拼写中非理性行为的表现。如综述部分所述，宏观层面非理性行为的普遍性已被多项研究证实，本文通过样本得到的结论与之相近，对此不再展开，下面着重对微观层面的非理性拼写行为进行分析。

微观层面的标签选用和拼写中的非理性行为有4个较为突出的特点。（1）普遍性，且该问题的普遍程度与用户的标签个数成正相关关系。以标签重复问题为例，样本中39.7%的用户存在该方面问题，且当用户标签数量少于10个时，出现该问题的概率为0，如图3所示。（2）不规范标签（如果用户的标签中有多个标签含义一致，则认为频次最高的标签为用户习惯使用的规范标签，其他为该用户的不规范标签）数量占总标签数量的比例较低，在存在该方面问题的用户中，不规范标签数量不足全部标签5%的占调查用户数量的75.5%，超过10%的仅占2.8%。（3）不规范标签的频次普遍较低，频次为1和2的占全部标签量的82.3%，少于5次的占94.3%。（4）不规范标签的类型比较集中，主要分布于中英文同义、国家/地区同义和类型/主题同义、繁简体方面。

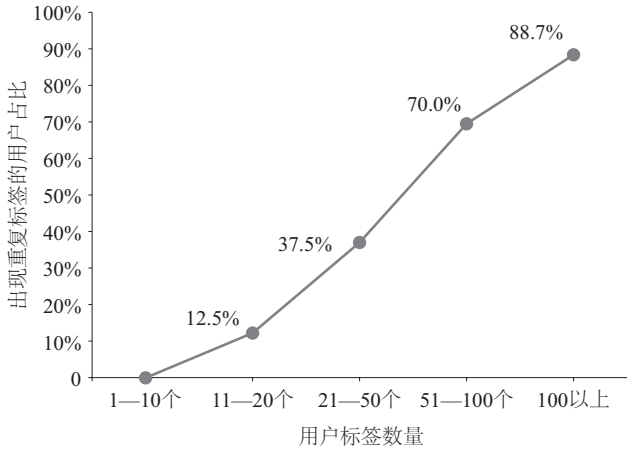

图 3 不同标签个数用户出现重复标签的比例

注：由于样本总体数量较少，采用等间距的统计方法会导致组间数据差别较大，但统计检验结果并不显著（卡方检验下，$p>0.05$）。基于此，采用间距不等的分组方式将分组临近且统计结果相近的组合并到一起。

通过以上分析可见，用户在社会化标注中存在较普遍的非理性标注行为，这些行为的表现形式多样，且涵盖社会化标注的全过程。

4 用户社会化标注中非理性行为产生的原因分析

基于15位豆瓣用户的访谈结果，导致用户社会化标注中非理性行为的主要原因可概括为用户态度、用户动机、用户标注习惯和社会化标注系统四方面。

4.1 用户态度的影响

在受访的15位用户中，13位用户对社会化标注的态度较为随意，均在标注中存在非理性行为。因而用户态度的随意性，是导致非理性标注行为最普遍且重要的原因。首先，在随意性态度的引导下，用户在选择标注对象时易受各种主客观因素的干扰，导致部分感兴趣的资源未被标注；其次，由于态度随意，用户通常不会在标注时按照一定的模式对资源特征系统分析，而是随兴而至，根据第一印象进行标注，从而导致在标注角度选择上缺乏统一性；最后，由于态度较随意，用户在标签拼写时不考虑是否规范、与之前的标签是否重复，也不会检查拼写是否有误，从而导致诸多拼写不规范、不一致的问题，有超过半数的受访者（8位）都存在这种情形。

4.2 用户动机的影响

不同类型的用户动机对非理性标注行为的影响不同，访谈结果显示，自我表达动机用户的非理性标注行为明显强于分类动机用户。首先，由于自我表达动机的用户只有在观影受到触动时才会标注，没有受到触动

或触动不够强烈时，不会进行标注，因此其更可能在标注对象选择上表现出不一致性。存在非理性标注行为的13位用户中，11位存在自我表达动机，7位表示其标注对象选择的非理性与此有关。其次，由于不同影片打动用户的角度不同，自我表达动机用户的标注角度更多样化，而分类动机用户则因分类维度较为固定，从而在标注角度上一致性更强。

4.3 标注习惯的影响

用户部分标注习惯也可能导致非理性标注行为增加，较为典型的是喜欢自己拼写标签的用户几乎都在标签选用和拼写方面存在非理性行为。其原因是用户对社会化标注的态度较为随意，因此在拼写时未考虑标签的规范形式或统一采用曾用过的表述方式，从而导致较多的同义词、简繁体、中英文等标签选用问题，也带来不少错误拼写问题。同时，在访谈中还发现13位存在非理性标注行为的用户中有5位会限制单个资源的标签数量，该习惯加剧了标注角度选择非理性。其原因是用户在确定标签的先后顺序上较为随意，通常由其即时念头决定，而标签的数量又受到限制，导致用户添加的标签只部分反映了其对资源的认知，客观上呈现出标注角度选择随意的特征。

4.4 社会化标注系统的影响

社会化标注系统主要从词汇控制、推荐标签类型及排序两方面引发用户的非理性标注行为，这也得到受访者的普遍认可。在词汇控制方面：一是在标签使用和拼写上不加限制，导致私人标签、同义标签、繁简体、英文复合词和单复数、形态等问题较严重；二是在标签推荐时缺乏词汇控制，导致标签中存在同义、不规范拼写，甚至个人化色彩较强的标签，加剧标签选用和拼写的非理性。在系统推荐的标签及其排序方面，由于对各资源推荐的标签类型不一致，且各类标签的排序差异较大，从而加剧用户标注角度选择非理性。

5 结论与启示

深入理解用户社会化标注中的非理性行为，有助于推动标签研究与应用的深入。本文基于用户日志和访谈数据，对社会化标注中非理性行为的表现和原因进行分析，认为用户社会化标注非理性行为主要表现在标注对象选择、标注角度和标签表达三方面。其原因与用户态度的随意性、标注动机、标注习惯及社会化标注系统有关。

基于上述结论作出进一步推断。（1）基于标签进行用户兴趣分析的结果可能是不可靠的，尤其在用户标签数量较少的情况下。基于标签进行用户兴趣分析，是标签较为流行的应用方式，但由于用户社会化标注行为的随意性，其标注的资源分布可能与其感兴趣的资源分布不同，且添加的标签可能没有反映其完整认知，因此基于标签进行用户兴趣分析的结果可能是不可靠的；且当用户标注的资源较少时，这种随意性对标注结果影响更大。（2）当参与标注的用户人数较少时，资源的高频标签未必能反映其最受用户关注的特征。在社会化标签的应用中，频次常被作为判断其与资源关系的重要特征，但由于用户社会化标注中非理性行为的大量存在，标签的频次高低并不能完全反映用户的真实认知，甚至也不能定性反映用户的真实认知，特别是在参与标注的用户人数较少时。

该结论对社会化标签研究与应用具有一定的启示意义。（1）在进行用户兴趣分析时，需要采用差异化策略，对于标签数据较为丰富的用户可采用基于用户标签的兴趣建模方法，但对于标签数据相对稀疏的用户，需采用其他方法进行分析。（2）在基于标签进行资源特征挖掘时，为提高结果准确率需设置参与标注的用户数量下限。

此外，研究样本的选择也可能导致结论具有一定局限性。本研究的日志数据采用的是“豆瓣电影”影视标注数据，其标注具有一定的延时性，即用户在完成影片观看一段时间后再进行标注，其行为可能与网页、在线视频网站等，即时性标注资源有所区别，导致结论的通用性不足。

参考文献

[1] MORRISON J P.Tagging and searching:search retrieval effectiveness of folksonomies on the World Wide Web[J].Information Processing &Management,2008,44(4):1562-1579.

[2] 吴克文,朱庆华,赵宇翔,等.社会化标注系统中标签检索质量模拟研究[J].情报学报,2011,30(1):29-36.

[3] SPITERI L F.Structure and form of folksonomy tags:the road to the public library catalog[J].Information Technology and Libraries,2013,

26(3):13-25.

[4] RAMAGE D,HEYMANN P,MANNING C D, et al.Clustering the tagged web[C]//Proceedings of the second ACM international conference on web search and data mining.New York:ACM,2009:54-63.

[5] 张富国.基于标签的个性化项目推荐系统研究综述[J].情报学报,2012,31(9):963-972.

[6] FIRAN C S,NEJDL W,PAIU R.The benefit of using tag-based profiles[C]//Latin American Web Congress.New York:IEEE,2007:32-41.

[7] 毛进,易明,操玉杰,等.一种基于用户标签网络的个性化推荐方法[J].情报学报,2012,31(1):23-30.

[8] 魏建良,朱庆华.社会化标注理论研究综述[J].中国图书馆学报,2009,35(6):88-96.

[9] NORUZI A.Folksonomies:(un)controlled vocabulary?[J].Knowledge Organization,2006,33(4):199-203.

[10] GU X,WANG X,LI R,et al.Measuring social tag confidence:is it a good or bad tag?[M]//CHEN L,TANG C J,YANG J,et al.Web-Age information management.Berlin:Springer Verlag GmbH,2011:94-105.

[11] BISCHOFF K,FIRAN C S,NEJDL W, et al.Can all tags be used for search?[C]//Proceedings of the 17th ACM conference on information and knowledge management.New York:ACM,2008:193-202.

[12] KIPP MEI,CAMPBELL D G.Patterns and inconsistencies in collaborative tagging systems:an examination of tagging practices[J]. Proceedings of the American Society for Information Science and Technology,2006,43(1):1-18.

[13] THOMAS M,CAUDLE D M,SCHMITZ C M.To tag or not to tag?[J]. Library Hi Tech,2009,27(3):411-434.

[14] 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,2010,30(6):76-81.

[15] 吴丹,林若楠,冯倩然,等.社会标签的规范性研究——图书标注[J].图书馆论坛,2012,32(1):1-7,56.

[16] 贾君枝,孙智超,郎杨芳.基于受控词表的医学资源社会化标签推荐研究[J].情报学报,2013,32(12):1326-1332.

[17] 熊回香,廖作芳,蔡青.典型标签本体模型的比较分析研究[J].情报学报,2011,30(5):479-486.

[18] 熊回香,邓敏,郭思源.国外社会化标注系统中标签与本体结合研究综述[J].情报杂志,2013(8):136-141.

[19] STROHMAIER M,KÖRNER C,KERN R.Understanding why users tag: a survey of tagging motivation literature and results from an empirical study[J].Web semantics: science, services and agents on the World Wide Web,2012(17):1-11.

[20] MIRZAEE V,IVERSON L.Tagging:behaviour and motivations[J]. Proceedings of the American Society for Information Science and Technology,2009,46(1):1-5.

[21] SEN S,LAM S K,RASHID A M,et al.Tagging,communities,vocabulary, evolution[C]//Proceedings of the 2006 20th anniversary conference on somputer supported cooperative work.New York:ACM,2006:181-190.

[22] GOLDER S A, HUBERMANB A.Usage patterns of collaborative tagging systems[J].Journal of Information Science,2006,32(2):198-208.

[23] BINKOWSKI P J.The effect of social proof on tag selection in social bookmarking applications[D].North Carolina:The University of North Carolina at Chapel Hill,2006.

[24] 豆瓣关注榜(此榜单仅收录5500以上关注度的豆瓣用户)[EB/OL].(2012-03-20)[2016-12-13].https://site.douban.com/144692/widget/forum/7144906/discussion/44924707/.

[25] 孙晓娥.深度访谈研究方法的实证分析[J].西安交通大学学报(社会科学版),2012(3):101-106.

作者简介

林鑫，男，1987年生，博士，华中师范大学信息管理学院讲师，E-mail：xinlin@mail.ccnu.edu.cn。
梁宇，女，1987年生，硕士，国网鄂州供电公司电力调度控制中心中级工程师，研究方向：电力信息分析。

Irrational Behavior in Social Tagging: Manifestations and Causes

LIN Xin[1], LIANG Yu[2]
(1.School of Information Management of Central China Normal University, Wuhan 430079, China;
2.Center for Electric Operations Control of Ezhou Power Supply Company, Ezhou 436000, China)

Abstract: In order to deepen the understanding of Users' social tagging behavior mechanism, and promote the development of social tags' application research and practice, this paper studies the manifestations and causes of irrational behavior in social tagging based on log file analysis and in-depth interview.The results show that, irrational behavior is very common in social tagging, which manifests in tagging objects choosing, tagging aspects choosing and tags expression; and the reasons are mainly related to users' attitude, tagging motivation, tagging habits and social tagging system.

Keywords: Social Tagging; Tag; User; Irrational Behavior

（收稿日期：2016-10-31）