社会化标注系统中不同主题资源的用户标注行为差异分析

武 强

（1 山西医科大学 太原 030001
2 中国人民解放军总医院《老年心脏病学杂志（英文版）》编辑部 北京 100853）

贺培凤 邰杨芳 卢学春 晋晓强 秦 勤

（山西医科大学 太原 030001） （中国人民解放军总医院 北京 100853） （太原师范学院 晋中 030619） （山西医科大学 太原 030001）

〔摘要〕 采用自主研发的软件抓取豆瓣网用户对互联网、健康、心理学等主题资源的标注标签数据，利用标签类型比率量化指标、统计描述法和差异分析法从7 方面分别进行系统性分析，结果显示不同主题资源在语言类型、标注倾向和词来源3 个方面对用户标注行为产生显著影响。

〔关键词〕 主题资源；标注行为；社会化标签；标签标注；差异分析
〔中图分类号〕R－056 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673－6036.2019.09.015

Difference Analysis on Users′ Tagging Behavior of Thematic Resources in Social Tagging System *WU Qiang, 1Shanxi Medical University, Taiyuan 030001, China, 2Editorial Office of Journal of Geriatric Cardiology, Chinese PLA General Hospital, Beijing 100853, China; HE Peifeng, TAI Yangfang, Shanxi Medical University, Taiyuan 030001, China; LU Xuechun, Chinese PLA General Hospital, Beijing 100853, China; JIN Xiaoqiang, Taiyuan Normal University, Jinzhong 030619, China; QIN Qin, Shanxi Medical University, Taiyuan 030001, China*

〔**Abstract**〕 The paper systematically analyzes Douban users´tagging data, collected by self－dependent software, concerning thematic resources like Internet, health and psychology by using quantitative indexes of tag type ratio, descriptive statistics and difference analysis in seven aspects respectively. The result shows that different thematic resources influences users′ tagging behavior as to three aspects of language types, the trend of tagging and the source of words.

〔**Keywords**〕 thematic resources; tagging behavior; social tags; tag tagging; difference analysis

〔修回日期〕 2019－06－13
〔作者简介〕 武强，硕士研究生；通讯作者：贺培凤，教授，硕士生导师，主持及参与各类科研项目20 项，主、参编教材5 部，发表论文120 余篇。
〔基金项目〕 国家社会科学基金项目“基于框架网络本体的标签系统语义分析研究”（项目编号：13TCQ030）；山西省科技基础条件平台项目“山西省医学科技文献共享服务平台”（项目编号：201605D121012）；2017 年度山西省软科学研究计划一般项目“社会化标注系统中的用户个性化医疗信息推荐研究”（项目编号：2017041036－1）。

1 引言

随着2002 年互联网开启 Web 2.0 技术主导的新时期大门，基于 Web 2.0 技术的互联网应用平台也纷纷涌现出来，社会化标注是 Web 2.0 技术的一种典型应用[1]。社会化标注在互联网、电子商务和图书馆等领域的重要性日益凸显，越来越多的互联网应用平台采用社会化标注功能，允许用户自由地发布、分享和评论网络信息资源，以关键词、字符、数字等标识的形式对网络信息资源自由地添加标签，方便用户描述并揭示网络信息资源的内容，为用户今后查找该资源提供便捷。

然而，社会化标注系统中的标签由用户自发定义且具有共享特点，在越来越多的社交分享网站（如 Delicious、豆瓣网、新浪微博、青棵网等）中标签存在选择随意、语义表达模糊、义同词不同以及词同义不同等问题，不但降低了用户对信息的使用效率，限制了标签的实际应用效果，而且也在一定程度上降低基于标签的信息组织和检索系统的质量[2]。面对社会化标注系统存在的问题，对用户标注行为展开系统性研究，尤其是基于标签对不同主题资源的用户标注行为进行差异分析，一方面有助于提高标签标注质量，另一方面可为优化社会化标注系统的功能和服务以及网络应用平台的研发与设计提供参考。

2 相关研究情况

2.1 概述

社会化标注技术在各大互联网平台上的推广应用为人们管理、检索和分享信息资源带来便利。同时也吸引国内外相关学者对社会化标注系统中基于标签的用户标注行为的关注与研究。国外学者研究起步较早，关注问题面较广，研究成果较为成熟，主要侧重标签形式、功能特征等方面；而国内学者却只侧重于对标签形式特征分析。相关研究可以归纳为标签形式与功能特征研究。

2.2 标签形式特征研究

通过对现有文献的调研与梳理，发现国外学者最早通过标签形式特征的揭示研究用户标注行为。如 Kipp 和 Cambell 从标签频率以及词汇共现频率等角度对 Delicious 标注系统中的标签词统计分析，发现标签的缩写、同义词、语言类型等种类形式繁多[3-4]。国内学者近年也在开展标签形式特征研究，如贾君枝从标签的长度、简称/缩写形式、频率、语法结构等视角对 Delicious 标注系统中标签进行特征分析[5]。

2.3 标签功能特征研究

国内外研究学者从划分标签类型和功能等视角研究用户标注行为。国外学者 Golder 和 Huberman 以 Delicious 的流行性标签集以 230 个用户为研究对象，统计分析发现用户的兴趣变化可以通过选择的标签得以体现，陈列出标签的功能类型[6]。Sen S 和 Lam K 等学者将标签划分为客观、主观和个人标签[7]。国内学者胡潜、石宇以计算机、心理学、经济学、文学作品和绘画 5 类图书用户数据构建标签分类体系，实证分析图书主题对用户标签使用行为影响[8]。梳理国内外相关研究，笔者发现目前关于社会化标注系统中基于标签的用户标注行为研究大部分只是用来揭示用户标注行为的特征、特点和一般规律；较少对其展开系统性研究，尤其对不同主题资源下用户标注行为的差异研究更少；同时缺乏相关的研究文献。

3 数据与研究方法

3.1 数据

3.1.1 来源 豆瓣网是当前用户参与度最高的社会化标注资源网站之一。该网站向图书和电影爱好者用户分别提供图书资源和电影资源的标签标注功能。本研究以豆瓣网上用户对图书资源标注的标签为数据，运用自主开发的标签抓取软件，选择“豆瓣读书”主页浏览区的互联网、健康、心理学等主

题资源，按照图书在豆瓣网上的热度排名分别抓取该类目下标注用户数≥30，且包含互联网、健康、心理学等标签的前500本图书的标注数据作为不同主题资源标签标注数据集，采集到的标注数据集里具体信息项包括：图书名、ISBN、图书url、标注图书状态和标注标签等信息。对各资源标注数据集的标签数据进行半自动清洗等处理，将处理后的标注数据频次出现大于10的标签作为本研究的数据对象。

3.1.2 处理 （1）标签数据清洗。利用自主开发的抓取软件工具中的清洗功能对不同主题资源的用户标注标签进行半自动清洗，包括去除无意义的字符、拆分复合标签，简化中文繁体字、转化英文大小写等。（2）标签的分词与词性统计。结合语料库在线平台[9]和自然语言处理与信息检索共享平台，运用分词功能将标签分为可切分词与不可切分词，结合使用Excel对不可切分词做词性统计。（3）标签词与受控词表的比较。将“情感分析用词语集”中的评价词及情感词、《中国分类主题词表》中的所有类目名称、主题词分别摘取至Excel表格中，同中文标签进行对比分析，非主题词采用深圳大学图书馆公共目录检索系统IVV1.0使用的入口词。

3.2 统计方法

3.2.1 统计描述法 即用少量数字（描述指标）概括大量原始数据，对数据进行描述的统计方法。本研究综合运用Excel、SPSS等统计软件对采集到并清洗后的不同主题资源的用户标注标签数据进行统计描述。

3.2.2 差异分析法 通过标签类型比率量化指标从标签的语言类型、词性统计、用语规范性、标注倾向、功能类型、情感、词来源7方面采用交叉列联表χ^2检验和双侧近似概率值P值判别不同主题资源对用户标注行为是否存在差异。

4 不同主题资源用户标注行为差异分析

4.1 概述

为全面考察不同主题资源对用户的标注行为是否存在显著差异，本研究采用用户标注行为量化指标——标签类型比率（TR），见表1，从标签的语言类型、词性统计、用语规范性、功能类型、情感、标注倾向、词来源等视角对互联网、健康、心理学等不同主题资源的标注标签数据进行差异分析。同时为检验不同主题资源间的差异是否具有科学的统计学意义选择卡方检验。卡方检验主要用于检验某无序分类变量各水平在两组或多组间的分布是否一致[10]。

表1 用户标注行为量化指标说明

指标	指标公式	意义
标签类型比率（TR）	$TR=\frac{\text{该类型的标签个数}}{\text{标签总个数}}$	描述标签类型的使用程度，是不同类型的标签个数与标签总个数之比

4.2 标签语言类型差异分析

标签语言类型反映用户对图书资源添加标签时选择语言的倾向程度。对用户使用标签语言类型频度信息的统计，见表2，用户在互联网、健康、心理学等不同主题资源的标注标签数据中，总体倾向使用中文标签标注资源。其中在健康主题资源的标注标签数据中表现的更加明显，中文标签在3个主题下的标签分类中所占百分比最高，TR高达95.9%。其次，互联网主题资源的标注标签数据中比起另外两个主题使用英文标签的TR比较偏高，说明互联网的用户倾向于使用专业英文术语标注资源。采用卡方检验方法判别不同主题资源对用户标注行为是否存在显著差异。检验结果显示Pearson值为108.551，$P=0.000<0.001$；说明在标签语言类型方面，不同主题资源对用户标注行为的差异具有统计学意义。为更好地解释这一差异，结果显示在中文标签标注中，互联网与健康资源、互联网和心理学资源两组组间比较都存在明显差异，健康和心理学资源组间却无明显差异；在英文标签标注中，互联网与健康资源、互联网和心理学资源、健康和心理学资源均存在明显差异；在数字标签中，互联网与健康资源、互联网和心理学资源、健康和心理学资源均无明显差异，见表3。

表 2 主题类型对标签语言类型的标注行为差异分析

主题类型		标签类型			χ^2 值	P 值
		中文标签	英文标签	数字标签		
互联网	n	930	154	14	108.551	<0.001
	TR（%）	84.7	14.0	1.3		
健康	n	971	29	12	–	–
	TR（%）	95.9	2.9	1.2		
心理学	n	1 404	84	17	–	–
	TR（%）	93.3	5.6	1.1		

表 3 主题类型对标签语言类型的标注行为组间比较分析

组间主题类型比较		标签类型			χ^2 值	P 值
		中文标签	英文标签	数字标签		
互联网	n	930	154	14	83.053	<0.001
	主题类型内（%）	84.7	14.0	1.3		
健康	n	971	29	12	–	–
	主题类型内（%）	95.9	2.9	1.2		
互联网	n	930	154	14	54.844	<0.001
	主题类型内（%）	84.7	14.0	1.3		
心理学	n	1 404	84	17	–	–
	主题类型内（%）	93.3	5.6	1.1		
健康	n	971	29	12	10.411	0.005
	主题类型内（%）	95.9	2.9	1.2		
心理学	n	1 404	84	17	–	–
	主题类型内（%）	93.3	5.6	1.1		

4.3 标签词性统计差异分析

标签词性反映用户对图书资源添加标签时选择词性的倾向程度。结合语料库在线平台和自然语言处理平台的分词功能，分别将互联网、健康和心理学等主题资源清洗后的中文标签进行分词。经统计，互联网、健康和心理学等不同主题资源的中文标签中可切分词分别有 612 个、606 个、984 个，不可切分词有 318 个、365 个、420 个。随后对不可切分词进行词性差异分析。卡方检验结果显示，Pearson 值为 14.382，$P = 0.422 > 0.05$；说明在标签词性统计方面，不同主题资源对用户标注行为的差异不具有统计学意义。分析其原因，可能是用户在社会化标注系统中标注资源时，用户个人自身不同文化程度的影响，对使用诸如习用语、缩略语、代词等词性标签标注资源的倾向程度不高。

4.4 标签用语规范性差异分析

将《中国分类主题词表》中的所有类目名称、主题词摘取至 Excel 表格中，分别将收集到的互联网主题资源中 1 098 个标签、健康主题资源中 1 012 个标签和心理学主题资源中的 1 505 个标签与《中国分类主题词表》中的类目名称及所有主题词，结合深圳大学图书馆公共目录检索系统 IVV1.0 进行比较分析，结果显示互联网主题资源的 930 个中文标签、健康主题资源的 971 个中文标签、心理学主题资源的 1 404 个中文标签中正式主题词占比分别为 23.33%、22.04%、22.51%；同分类类目名称比对分析，结果显示中文标签中有类目名称占比分别 4.84%、6.59%、7.05%。同时标签词中分别存

在1.72%、2.78%和1.35%的入口词。标签用语规范性是指用户使用标签标注资源时标签用语是否具有规范性，在一定程度上体现社会化标注系统中用户标签选择的个性化和随意性。类似分析结果显示，分类型 Pearson 值为 4.850，$P = 0.088 > 0.05$；功能型 Pearson 值为 6.889，$P = 0.142 > 0.05$；均说明在标签用语规范性方面，不同主题资源对用户标注行为的差异不具有统计学意义。众多参与社会化标注系统标注图书资源的用户主要为获取知识，对标注资源所采用标签用语不太注意是否规范。

4.5 标签功能类型差异分析

标签功能类型主要反映用户是否选择受控词表中的分类类目名称、正式叙词作为标签，在一定程度上不仅体现用户使用标签组织管理图书资源的便捷性，而且也体现用户揭示图书主题内容的倾向性。类似分析结果显示，分类型 Pearson 值为 4.850，$P = 0.088 > 0.05$；功能型 Pearson 值为 6.889，$P = 0.142 > 0.05$；均说明在标签功能类型方面，不同主题资源对用户标注行为的差异不具有统计学意义。分析原因，用户在社会化标注系统中标注资源时，由于自身所储备的知识有限以及大部分用户并非信息组织和管理的专业人士，对标签是否为正式主题词和类目名称等信息并不了解，以致于标签所具备的分类和描述功能往往被忽略。

4.6 标签情感差异分析

将清洗后的不同主题资源中文标签分别同知网发布的“情感分析用词语集（beta 版）”（http://www.keenage.com/download/sentiment.rar）进行对比，统计发现互联网、健康、心理学等不同主题资源中文标签分别有评价型标签 17 个、18 个和 34 个；情感型标签 3 个、3 个和 8 个。标签情感主要包括对图书本身内容信息客观描述的评价型标签、表达用户对图书主观感受的情感型标签和其他标签。而评价型和情感型标签在一定程度上便于其他用户关注和了解资源，有利于资源共享。类似分析结果显示，Pearson 值为 2.597，$P = 0.627 > 0.05$，说明在标签情感方面，不同主题资源对用户标注行为的差异不具有统计学意义。分析原因，可能是众多用户在图书阅读标注过程中，只凭借自身判断力对图书简介部分中精彩的内容信息做出客观描述和主观感受；同时，也存在用户阅读一些社会知名度较高作者的图书作品。这些因素影响用户标签情感的表达，致使不同主题资源类型对用户标注行为的影响差异不明显。

4.7 标签标注倾向差异分析

标签标注倾向可通过用户对图书资源的标注率体现出来，通过测量用户已标注图书与未标注图书的标签使用频度来反映。类似分析结果显示，至少有 57% 用户从未对图书添加标签；17% 用户会为自己阅读的所有图书都添加标签；11.5% 用户时而添加，时而不添加。同时，Pearson 值为 32.713，$P = 0.036 < 0.05$，说明在标签标注倾向方面，不同主题资源对用户标注行为的差异具有统计学意义。通过主题类型对标签标注倾向的标注行为组间比较分析，结果显示在标签标注倾向的标注行为上，互联网和健康资源存在明显差异，互联网和心理学资源、健康和心理学资源均不存在明显差异。利用这一差异，社会化标注系统应不断完善和引导用户充分利用标签标注功能，提高用户对图书资源的标注率。也说明标签标注倾向对社会化标注系统的组织和网络信息资源的管理均可发挥重要作用。

4.8 标签词来源差异分析

是指用户在确定资源信息时对标签选择倾向性和偏爱性大小，在一定程度上反映用户标注资源的积极性。标注资源的标签主要来源于其他用户自定义的标签、系统推荐的标签和资源标题提取的标签。通过分别对收集到的互联网主题资源 500 本图书 1 098 个标签、健康主题资源 500 本图书 1 012 个标签与心理学主题资源 500 本图书 1 505 个标签进行统计分析，发现分别有 988 个、883 和1 369个标签并非来源于标题，占标签总数的 89.98%、87.25% 和 90.96%。类似分析结果显示，3 个不同主题资源的标签数据集中至少有 9.0% 中文标签来自于标题；同时，Pearson 值为 9.175，$P = 0.010 <$

0.05；说明在标签词来源方面，不同主题资源对用户标注行为的差异具有统计学意义。通过主题类型对标签词来源的标注行为组间比较分析，结果显示在标签词是否来源于标题的标注行为上，互联网和健康资源、健康和心理学资源都存在着明显差异，然而互联网和心理学资源却不存在差异。利用这一差异，社会化标注系统应引导用户充分利用标签个性化推荐功能，将更多的网络信息资源分享和利用。

5 结论与建议

5.1 结论

本研究采用统计描述法与差异分析法，以标签类型比率量化指标对豆瓣网中的互联网、健康和心理学3 大不同主题资源，从语言类型、词性统计、用语规范性、标注倾向、功能类型、情感、词来源等7 方面做系统性地分析，发现不同主题资源对用户标注行为具有显著性差异，体现在标签的语言类型、标注倾向和词来源等方面，在标签的词性统计、用语规范性、功能类型和情感等方面，不同主题资源对用户标注行为影响差异不明显。

5.2 建议

社会化标注系统应利用不同主题资源对用户标注行为差异的特性，从提高个性化推荐服务功能、增加标签导航功能和完善标签检索功能等方面优化社会化标注系统的相应功能和服务，以此促进社会化标注系统的可持续发展。

参考文献

1 Marelow C Naaman M, Boyd D, etal. Position Paper, Tagging paper, Taxonomy, Flickr, Article, To Read [C]. New York: Proceedings of the 17th Conference on Hypertext and Hypermedia, 2006.

2 李蕾，王冕，章成志. 区分标签类型的社会化标签质量测评研究 [J]. 图书情报工作，2013，57（23）：11－16.

3 Kipp M E I, Campbell DG. Patterns and Inconsistencies in Collaborative Tagging Systems: an examination of tagging practices [J]. Proceedings of the Amenican Aociety for Iuformation & Science and Techology, 2006, 43（1）: 1－18.

4 Spiteri LF. The Structure and Form of Folksonomy Tags: the road to the public library catalog [J]. Information Technology and Libraries, 2007, 26（3）: 13－24.

5 贾君枝，王东元，王永芳. 基于Delicious 中文标签特征分析 [J]. 情报科学，2010，28（10）：1565－1568.

6 Golder SA. Usage Patterns of Collaborative Tagging Systems [J]. Jounnal of Information Science, 2006, 32（2）: 198－208.

7 Sen S, Lam SK, Rashid AM, et al. Tagging, Communities, Vocabulary, Evolution [C]. USA: Proceedings of the Conference on Computer Supported Cooperative Work, 2006: 181－190.

8 胡潜，石宇. 图书主题对用户标签使用行为影响研究 [J]. 图书情报工作，2016，60（8）：106－112.

9 北京汉语国际推广中心，北京师范大学中文信息处理研究所 [EB/OL]. [2018－10－10]. http://www.aihanyu.org/cncorpus/index.aspx.

10 陈景祥. R 软件应用统计方法 [M]. 大连：东北财经大学出版社，2014：396.