中国期刊方阵“双奖”期刊
国家期刊奖提名奖期刊
全国中文核心期刊
RCCSE中国权威学术期刊（A+）
CSSCI来源期刊
中国优秀图书馆学期刊

ISSN 0252-3116
CN11-1541/G2

图书情报工作®

LIBRARY AND INFORMATION SERVICE

本期要目

2013 23 VOL. 57

● 中国科学院 主管 ● 中国科学院文献情报中心 主办 ●《图书情报工作》杂志社 出版

Tushu Qingbao Gongzuo
图书情报工作
LIBRARY AND INFORMATION SERVICE

国家级大型图书馆学情报学核心期刊
全国优秀科技期刊

半月刊　每月5日、20日出版
2013年12月　第57卷　第23期
总第57卷第492期　1956年创刊

主　管　中国科学院
主　办　中国科学院文献情报中心
协　办　中国图书馆学会专业图书馆分会
出　版　《图书情报工作》杂志社
地　址　北京中关村北四环西路33号（100190）
电　话　（010）82623933
　　（010）82626611-6614
传　真　（010）82621460
网　址　http://www.lis.ac.cn
邮　箱　journal@mail.las.ac.cn
　　tsqbgz@vip.163.com
微　博　weibo.com/lis1956

社长、主编　初景利
副 主 编　易　飞
编辑部主任　杜杏叶
编 辑 部　易　飞　杜杏叶　徐　健　王传清　王善军　刘远颖　赵　芳　谢梦竹
英文编辑　魏　蕊　胡　芳
本期责编　易　飞

印刷单位　北京科信印刷有限公司
国内总发行　北京报刊发行局
国外总发行　中国国际图书贸易总公司
订　购　处　全国各邮电局
广告经营许可证　京海工商广字第0032号
广 告 审 核　苗志刚
刊号　ISSN 0252-3116　CN11-1541/G2
定价　68.00元
国内邮发代号　2-412
国外邮发代号　M215

获奖及被收录情况

中国期刊方阵"双奖"期刊
国家期刊奖提名奖期刊
全国中文核心期刊
RCCSE中国权威学术期刊（A+）
CSSCI来源期刊
中国优秀图书馆学期刊

本刊被以下数据库收录：
LISA（图书馆和信息科学文摘）
中文社会科学引文索引（南京大学）
中国期刊全文数据库（清华同方）

版权所有©《图书情报工作》杂志社，未经许可不得转载

CONTENTS 目录

学术前沿

专题：Web2.0上社会化标签的深度挖掘

理论研究

工作研究

中英译本图书社会化标签的比较研究*

卢超 章成志

[摘 要] 认为随着 Web2.0 的发展,社会化标注系统也开始影响图书馆传统的信息组织方式。以中英译本的图书为研究对象,从豆瓣、Amazon、Librarything 获取社会化标签,从中国国家图书馆和美国国会图书馆获取 MARC 记录中的主题词,探究每本书的标签和主题词在长度、个数和相似度三方面的差异。实验结果显示:不同语种的标签或主题词在长度、个数和相似度三方面存在差异;不同类目的图书的社会化标签在这三方面也存在差异;单本书的社会化标签与主题词之间同样存在差异。该研究能够帮助图书馆了解社会化标签并借此提升用户服务品质。

[关键词] 社会化标签 图书标引 主题词 标注系统
[分类号] G350
DOI:10.7536/j.issn.0252-3116.2013.23.003

Web2.0 的兴起让用户从网络信息的接收者转变成网络信息的提供者和组织者。信息组织是指将处于无序状态的特定信息,根据一定的原则和方法,使其成为有序状态的过程[1]。用户在 Web2.0 的环境下可按照自己的认知、喜好与习惯组织各种网络信息资源。而社会标注(social annotation,或 social tagging),作为 Web2.0 环境下的一种典型应用被广泛用来进行新环境下的信息资源组织。社会标签则是其组织资源的工具。

由于标签具备高分享性、简洁性、易用性等优点,社会标注开始进入图书标引领域,并得到用户的广泛运用。Lib2.0、OPAC2.0 等概念也相继被提出[2]。在 OPAC2.0 技术的构想和发展中,美国的一些图书馆利用社会标签标引图书,形成云图,辅助用户检索;或者将有价值的标签直接写入机读目录的 653 字段或 65X 中,帮助检索[3]。

社会化标注系统在信息组织和信息检索方面的巨大潜力也吸引了大量学者投入到社会标签的相关研究中来。这些研究包含探讨社会化标签的自身的特征以及与受控系统(如主题词)的比较两方面。然而这些研究往往关注标签的整体状况,缺乏对单本图书标注的研究;同时,社会化标签的比较研究也缺乏较为细致的跨语言研究。

为此,本文首先利用豆瓣、Amazon、Librarything 三大书评网站获取 1 200 本图书的社会化标签,并从中国国家图书馆(以下简称"国图")、美国国会图书馆获取这 1 200 本图书的 MARC 记录中的主题词,接着从标签和主题词的长度、个数以及相似度三个方面组织数据分析实验。根据实验结果探讨在不同语言、不同图书类目下社会化标签的长度、个数以及重合度方面的特点和差异,并与主题词进行比较。以上研究可以为以后的语义研究奠定基础,为建设 Web2.0 下的书评网站或者推进 Lib2.0 服务提供参考。

1 相关研究

1.1 社会化标签

社会化标签与博客、微博等应用一起推动着用户意识的转变。社会化标签主要承担网络资源的组织与分享,成为元数据的最佳来源。E. Tonkin 将标签描述为"任何能够描述网络上的资源和用户想法之间关系的词"[4]。

社会化标签具有其独特之处:①非受控性;②易用性;③共享性;④类聚性;除此以外,还有很多的性质如多样性、个性化等。

因为标签存在以上特性,社会标注系统中存在大

*本文系教育部人文社会科学基金规划项目"多语言高质量社会化标签生成及聚类研究"(项目编号:13YJA870020)和中央高校基本科研业务费专项资金项目"多语言高质量社会化标签生成及聚类研究"(项目编号:30920130132013)研究成果之一。
[作者简介] 卢超,南京理工大学经济管理学院硕士研究生;章成志,南京理工大学经济管理学院副研究员,博士,博士生导师,通讯作者,E-mail:zhangcz@njust.edu.cn。
收稿日期:2013-10-08 修回日期:2013-11-08 本文起止页码:17-23 本文责任编辑:杜杏叶

量低质量标签。抽取标签,提高标签质量显得尤为重要。目前生成标签的方式有:①基于标签推荐生成标签。A. Rae[5]、Xu Zhichen[6]、S. O. K. Lee[7]等依据用户的标签行为和用户间的关系生成标签并形成标签推荐。钟青燕等[8]、刘知远[9]进一步利用语义模型优化标签推荐。②基于文本抽取的生成方式。C. Brooks[10]、刘知远[9]等分别对博文和微博实现了内容类聚,形成标签;刘知远在此基础上还实现了文本的可视化。S. Shilad 等依据标签出现的频率试图发现高质量标签[11]。

1.2　社会标签与主题词比较研究

P. J. Rolla 通过对比主题词和社会化标签,发现社会化标签对图书的标注更加全面细致,能够提高检索性能,而主题词表却只能做基本信息的标引。因此,联合标签和主题词标引图书是一个不错的选择[12]。M. Thomas 等人的相关研究得到了同样的结论[13]。J. Park 等人总结前人研究,发现社会标注在图书标引方面的巨大潜力,但其中与主题词无关的标签将很难运用到传统的图书馆标引系统中[14]。吴丹以主题词作为参考系研究标签规范度,结论显示社会标签的规范度较低[15-16]。从该研究中可以看出,社会化标签和主题词在外形上存在差异。

目前,学术界对社会化标签在标注方面做了很多研究,主要从社会化标签的自身研究和与受控语词对比研究两个维度展开,涉及到不同语种、不同站点和不同图书的标注。但这些研究存在以下问题:①没有涉及大量的样本数据分析;研究的标注系统较少。②一般只关注某个领域或者某一类别的图书标注结果,并不涉及多个领域图书标注的比较分析。③多从标签的集合来研究标签与主题词的差别,并没有从单本书的角度对社会化标签进行细致研究。④缺少对多语言和单语言标注结果的比较研究。本文基于以上研究的不足探讨社会化标签在图书标引中的应用,为推进图书馆的服务提供参考。

2　实验数据采集与处理

本文利用几大标签系统站点以及图书馆 MARC 主题词获得图书标签数据,据此获得图书标签的平均长度、个数以及各站点数据之间的重合度。

2.1　实验数据采集

2.1.1　*实验数据采集概述*　在采集实验数据时,笔者对以下几点进行了控制:①图书获取。豆瓣读书根据标签的热点将图书分为六大类目。本研究按照这种分类标准,在每个类目下随机抽取 200 本图书,共计图书 1 200 本。②数据来源。本文所采集的标签和 MARC 主题词均来自豆瓣读书、国图、美国国会图书馆、Amazon 和 Librarything,当某本图书在国图没有 MARC 记录时,允许利用上海图书馆 MARC 主题词替代国图 MARC 主题词。为了控制标签的数量和质量,在 Librarything 上只采集 show number 下的默认标签。③数据有效性。一方面,同语言下图书信息需完全一致;不同语言之间,图书只存在译本的差异;另一方面,每条记录必须包含所有要获得的字段值,否则认定该记录无效。④采集时间。本文研究数据采集的时间段为 2013 年 4 月 1 日 00:00 - 4 月 15 日 24:00。

2.1.2　*实验数据采集内容*　利用从豆瓣获取的图书题名进行检索,找到五大网站上的相关数据。一条有效的图书标签记录必须包含:中文题名、英文题名、豆瓣标签及链接、中国国家图书馆 MARC 主题词及链接、美国国会图书馆 MARC 主题词及链接、Amazon 标签及链接和 Librarything 标签及链接。

2.2　实验数据预处理

本文实验前需要对数据进行标准化,具体的数据预处理包括标签长度、个数、重合度几个方面。

2.2.1　*标签长度、个数实验数据预处理*　每条记录中,每个标签和每个主题词之间都用“*”相隔。其中,计算标签和主题词长度时,汉字字符长 2 字节,其他字符长 1 字节,分别计算 5 个站点 1 200 条记录的长度和个数,统计结果。

2.2.2　*标签重合度实验数据预处理*　重合度实验前,将标签按频次降序排列,同频次按字母顺序升序排列;将 MARC 主题词中的复合词进行切割去重;再将标签和主题词标准化。计算重合度时,本文将标签按照排名依次取前 5、前 10、前 15、前 20 和全部标签共 5 层分别计算重合度。计算重合度有很多的计算方法,比如向量余弦值、D 系数、TF * IDF、以及 Jaccard 指标等其他很多方法。由于本文所采集的数据是以字段的形式存储的,同时考虑计算过程的简便性,本文最终选择 Jaccard 指标(Jaccard Index)计算每一本图书各站点间的重合度。其中 Jaccard 指标的计算公式如下:

$$\text{Jaccard}(X,Y)=\frac{|X\cap Y|}{|X\cup Y|} \tag{1}$$

其中,X,Y 分别表示某条记录中两标签集合;$|X\cap Y|$ 表示两集合共同标签个数,$|X\cup Y|$ 表示两集合所有互异标签数,二者的比值即为本文重合度。如:某条记录中豆瓣标签和中国国家图书馆的 MARC 主题词分别

为:“投资 * 股票 * 彼得林奇 * 金融 * 理财 * 经济 * 彼得林奇 * 战胜华尔街”和“股票 * 证券投资 * 经验 * 美国”,则二者全部标签的重合度为:

$$\text{Jaccard}(\text{豆瓣}, \text{国图}) = \frac{|X \cap Y|}{|X \cup Y|}$$

$$= 1/(12-1) = 0.09091$$

总之,本文需要计算豆瓣 - 中国国家图书馆、美国国会图书馆 - Amazon、美国国会图书馆 - Librarything、Amazon - Librarything 4 组重合度,每组重合度包括整体重合度和分层重合度,并为每组重合度获取相关统计量。举例:若豆瓣和中国国家图书馆进行重合度实验,某本书只有 8 个豆瓣标签、3 个主题词,则只计算整体重合度;若豆瓣标签为 8 个,主题词为 10 个,则除了计算整体重合度以外,还要计算一次 Top5 的分层重合度。其他情况以此类推。

3 结果分析

本文得到了标签长度、标签个数以及标签重合度的实验结果。具体结果分析如下:

3.1 外部特征分析

3.1.1 标签长度分析

- 整体情况分析。本文对实验数据进行统计分析,结果如表 1 所示:

表 1 各站点标签长度相关数据机器处理结果

项目 \ 站点	豆瓣	中国国家图书馆	美国国会图书馆	Amazon	Librarything
最小值	3	4	4	4.714 286	6
中值	6.25	15	23	11.333 33	9.066 667
众数	5.75	15	20	11	8.633 333
最大值	11.75	62	90	92	15.846 15
平均值	6.392 673	16.264 85	24.844 36	11.949 5	9.142 331
标准差	1.144 165	7.208 073	10.742 41	3.991 607	1.130 893

从表 1 提供的数据来看,豆瓣标签的长度一般为 5 - 7 个字节,即为 2 - 3 个汉字长度,且标准差较小,由此可见豆瓣标签多以短词为主;中国国家图书馆的主题词一般为 15 个字节,即主题词的长度总体在为 6 - 7 个汉字,并包含一些“ - - - ”字符,由于标准差较大,长度值波动也较大;美国国会图书馆的英文字符长度总体在 20 - 25 字节之间,但由于标准差最大,所以主题词长度波动最大;Amazon 的标签长度一般在 11 - 12 个字节之间,最大值为 92 字节,差距较大;但由于标准差较小,可见异常值并不多,整体较为稳定;Librarything 的标签长度一般在 8 - 9 个字节之间,标准差和豆瓣相似,表明 Librarything 的标签长度也较为稳定。

- 类别比较分析。根据豆瓣读书提供图书的分类信息,实验数据可分为 6 类,得到实验结果,如图 1 所示:

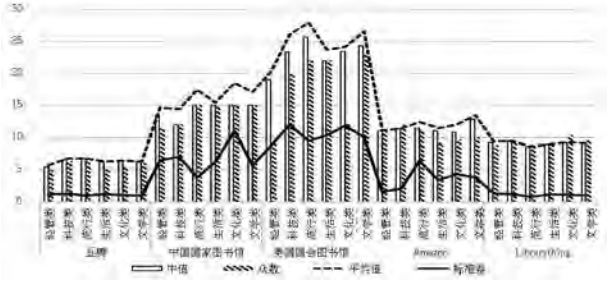

图 1 标签长度站点 - 类目组合

总体来看,豆瓣和 Librarything 6 个类目的长度的平均值和标准差的曲线都比较水平,波动并不大。中国国家图书馆和美国国会图书馆 6 个类目之间长度差异在均值和标准差方面都比较明显,但中国国家图书馆 6 个类目在均值方面的差异比美国国会图书馆大,而美国国会图书馆 6 个类目的标准差之间的差异要明显小于中国国家图书馆。Amazon 6 个类目的平均长度都比较接近,但标准差有较大差异。

3.1.2 标签个数分析

- 整体情况分析。相似地,标签个数统计结果如表 2 所示:

表 2 各站点标签数量相关数据机器处理结果

项目 \ 站点	豆瓣	中国国家图书馆	美国国会图书馆	Amazon	Librarything
最小值	1	1	1	1	1
中值	8	2	3	12	30
众数	8	2	2	12	30
最大值	13	7	14	17	365
平均值	7.912 5	2.138 333	2.992 5	10.26	29.33
标准差	0.709 646	1.023 101	1.914 985	3.124 719	12.974 9

表 2 显示,豆瓣的标签一般为 8 个,且标准差较小,可见豆瓣标签的数量形成了较为稳定的状态;中国国家图书馆和美国国会图书馆主题词的数量大体在 2 - 3 个,主题词数量的波动较小,其最大值分别为 7 和 14,显示存在主题词数量较多的个案;Amazon 的标签数一般在 10 - 12 个,标准差较大,但整体比较稳定;Librarything 的标签一般为 30 个,然而,其最大值为 365,最小值为 1,标准差也较大,表明 Librarything 标签数量存在不稳定因素。但需要注意的是豆瓣和 Librarything 所提供的标签个数都不是全部的用户标签,而是用户利用频次最高的前 8 或前 30 的标签。

从标准差显示的结果来看,标签系统中豆瓣的标

准差最小，其次是 Amazon，最大的是 Librarything。标签的平均个数可能也起到一定的放大作用。

• 类别比较分析。对各个站点标签和 MARC 主题词按照类目的分析方法为：先对每个站点的类目进行分析，然后将所有的结果汇总得到整体的组合图，具体如图 2 所示：

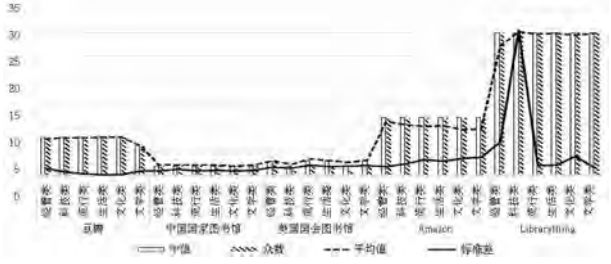

图 2　标签个数站点－类目组合

从图 2 来看，五大站点标签个数的均值、众数和中值都非常接近，并且表现在标准差方面都比较低，每个类的标签个数在各个站点的表现都比较均匀。但具体到每一个站点，标准差也有明显的峰值出现，即 Librarything 里的科技类标准差的变化十分明显。因此，在标签个数方面，五大站点的数据都比较平稳，没有太大的变化，除了 Librarything 中科技类的标准差有巨大的差异外。

3.2　语义特征分析

3.2.1　标签整体重合度分析　本实验中，本文按照语言分类检测重合度，得到 4 组：豆瓣－国图、美国国会图书馆－Amazon、美国国会图书馆－Librarything 以及 Amazon－Librarything。实验中，按照标签的标引频次划成：Top5、Top10、Top15 以及 All4 层测量，主题词取同权重。实验结果见表 3。

表 3 显示了各站点所有标签重合度按照区间对 1 200条数据进行统计的结果：

表 3　各站点标签整体重合度分布情况

数值 / 组别 / 区间	豆瓣－国图		美国国会图书馆－Amazon		美国国会图书馆－Librarything		Amazon－Librarything	
	平均值	计数	平均值	计数	平均值	计数	平均值	计数
0	0	432	0	567	0	194	0	74
(0,0.1]	0.095 51	517	0.069 56	425	0.053 68	875	0.060 94	329
(0.1,0.2]	0.141 52	207	0.142 02	175	0.133 46	131	0.145 16	636
(0.2,0.3]	0.231 31	41	0.238 86	25	0	0	0.238 64	141
(0.3,0.4]	0.347 22	3	0.317 95	5	0	0	0.321 49	20
(0.4,1]	0	0	0.555 56	3	0	0	0	0
合计	0.074 33	1200	0.053 03	1200	0.053 71	1 200	0.127 04	1 200

从表 3 可以看出，中文标签重合度多集中在[0，0.1]之间，约占总数的 79.1%，平均重合度为 0.052 03；其中重合度为 0 的共 432 个，占该区间的 45.5%，而另外的 517 条数据落在[0.05，0.1]之间，平均重合度为0.095 51。重合度最高的为 0.375，整体的平均值为0.074 33。

总体标签相似度最高的站点为 Amazon 和 Librarything。这两者标签的重合度平均表现最好，且重合度分布的区间也相对其他实验组更加稳定和均匀。总体表现最差的实验组为美国国会图书馆和 Amazon 之间的重合度数据。这两者之间的平均重合度为0.053 03，为 4 组数据中均值最低的一组；重合度为 0 的数据高达 567 个，占总体的 47.25%。

此外，笔者认为：①4 组的重合度数据除了 Librarything－Amazon 组都在[0，0.1]之间。这表明，各站点之间标签的总体相似度并不高。②重合度值为 0 的数据都比较多，这些特殊值的大量存在都会严重影响标签相似度的整体表现。③标签和主题词之间的相似度十分不理想，这从前三组实验和最后一组实验的数据比较可以看出来。

3.2.2　分层标签重合度分析　在分层标签的重合度实验中，本研究按照标签标引的频次将标签降序排列，依据实验组中最小标签数进行 Top5、Top10、Top15、Top20 四个层次的重合度分析。如此可以利用标引频次对高质量的标签进行比较，改善实验数据的表现。

豆瓣和国图实验组只存在 Top5 的标签实验。实验显示，有效数据为 62 个，重合度平均值为 0.065 41；其中重合度为 0 的有 34 个，占 54%。

美国国会图书馆和 Amazon 实验组存在 Top5 和 Top10 的实验结果。Top5 标签比较中，有效数据共 336 条，平均重合度为 0.065 4；其中重合度为 0 的有 232 个，占 69%。Top10 标签比较中，有效数据为 47 条，平均重合度为 0.043 42；其中重合度为 0 的有 16 个，占 34%。

美国国会图书馆和 Librarything 的重合度的实验结果显示，Top5、Top10、Top15 及 Top20 的重合度的平均值分别为 0.089 48、0.083 59、0.085 57、0.081 08；重合度为 0 的个数分别为 124、8、0、0，分别占 33%、15%、0%、0%。

总体标签相似度最高的站点为 Amazon 和 Librarything。这两者标签的重合度平均表现最好，且重合度分布的区间相对其他实验组也更加稳定和均匀。总体表现最差的实验组为美国国会图书馆和 Amazon 之间的重合度数据。这两者之间的平均重合度为0.053 03，为 4 组数据中均值最低的一组；重合度为 0 的数据高达 7.25%。

Amazon 与 Librarything 标签重合度实验指出，

Top5、Top10 的重合度的平均值分别为 0.190 32、0.0.194 75;重合度为 0 的个数分别为 168、36,分别占 16%、4%。

笔者将得到的各层标签重合度的均值制成图,在某组数据缺损时,利用整体重合度的数据填充,结果如图 3 所示:

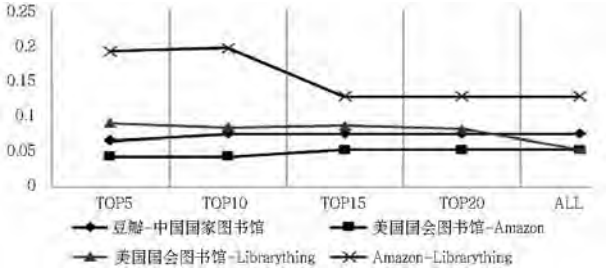

图 3 各组重合度均值折线

从图 3 看,不管是分层还是整体的重合度,Amazon 和 Librarything 的表现都是最好的,且在每个阶段的表现均最佳;美国国会图书馆与 Amazon 标签的重合度的平均表现仍是最差的。纵观全图,一般标签的数量越少,重合度的表现也越好,这和标签的标引频次有关,即标引的频次越高,标签的质量也相对越高。

3.2.3 不同类别的标签重合度比较分析 整体重合度分析和分层重合度分析之后,将标签重合度再进行类目分析。类目分析按照两个方面展开:统计 6 个类目下各组重合度实验的有效数据个数;统计 6 个类目下各组重合度实验的平均值。有效数据用来反映重合度实验中两站点标签的最小个数状况;均值用来反映重合度在数值上的整体表现。

- 有效数据分析。图 4 展示了各站点 6 类目有效数据的统计情况。豆瓣和国图的有效数据仅出现在 Top5 和 All 两层,且 Top5 层只有文学类有效数据超过总数的 10%,其他类目数据很少;美国国会图书馆和 Amazon 重合度数据出现在 Top5、Top10 和 All 三层,Top5 层流行类有效数据最多,科技类最少,Top10 层数据量明显减少,最多的仍是流行类为 13 个;美国国会图书馆和 Librarything 的重合度有效数据表中,经管类的数据出现在所有层级中,其他类目都还是集中在 Top5、Top10 和 All 三层,有效数据最多的是流行类,其次是生活类;Amazon 和 Librarything 的统计表中,数据集中在 Top5、Top10 以及 All 三层,且 6 个类目的数据量在 Top5 中均超过总数的 85%,Top10 的有效数据也超过 60%,有效数据在各个类目之间的差异并不明显。
- 重合度均值分析。将得到的有效数据进行汇总计算得到 4 组重合度实验 6 个类目的均值,按照实验组别记录(见图 5):豆瓣和中国国家图书馆的重合度实验中,Top5 重合度均值最高的是文化类,最低的是科技类和流行类;整体重合度实验中,流行类的重合度最高,科技类最低。美国国会图书馆和 Amazon 的重合度实验中,Top5 重合度均值最高的是科技类,最低的是流行类;Top10 重合度均值最高的是流行类,最低的是生活类;整体重合度,最高的是科技类,最低的是生活类。美国国会图书馆与 Librarything 的重合度实验中,Top5 均值最高的是流行类,最低的是经管类;Top10 均值最高的是流行类,科技类最低;Top15 只有三组数据,最高的是生活类和文化类;整体重合度,最高的是生活类,经管类最低。

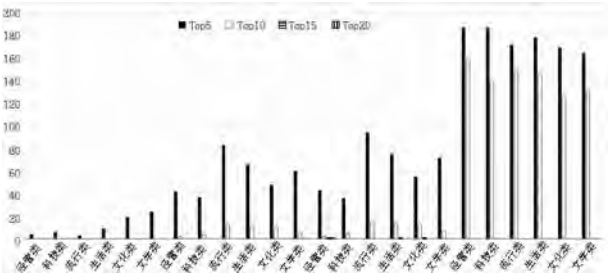

图 4 各组重合度实验有效数据类目分布

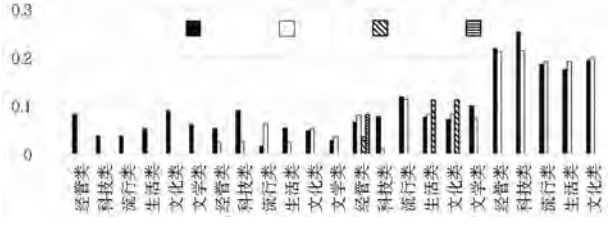

图 5 各组重合度实验均值类目分布

4 讨 论

4.1 实验结论探讨

4.1.1 外部特征探讨 标注系统中的标签长度短且标准差小,这充分体现了社会化标签简洁的特性。不同的站点间又存在差异,Amazon 长度的标准差相较其他两个系统会大很多,这可能是因为 Amazon 本质上仍是一个网上购物商,标签用户的同质性差。

不同类目下,流行类和文学类的图书的长度标准差都更小,这可能因为这两类标签之间的相似度更大;而科技类标签的长度时最小的,可能因为科技类标签存在更多的缩写词。MARC 主题词的长度比较长且标准差大,这和复合主题词的存在有直接的关系。社会化标签个数在不同语言、不同类目、不同站点中都表现出极大的相似性,标签的数量都受到严格控制;Amazon 的标签数量在类目中并不稳定,这可能因为用户数量

不足。虽然豆瓣、Librarything 的数据都不是全部标签，但这恰好说明这些社会化标注系统对标签的质量有一定的控制。然而在数据采集时，笔者仍然发现 Librarything 中不点击 show all 按钮仍存在大量的垃圾标签。MARC 主题词数量很少，一般在 2－3 个左右，主题词个数的标准差也明显比社会化标签小很多。从这一点来看，主题词标注图书有其明显的优势，但对网络资源的揭示程度有限。

总之，社会化标签数量多、长度小，而 MARC 主题词数量少、长度可能很大。复合词主题词的存在，使得主题词的数量维持在比较低的水平，但长度的增加必然增加了标注的难度和用户识别的难度。社会化标签长度短、类型多、便于用户使用和识别，又能多方位揭示资源。所以，从外部特征来看，社会化标签的一些优点恰好弥补了 MARC 主题词存在的不足。

4.1.2　语义特征探讨　重合度分析时，笔者发现：不同语言下标签与 MARC 主题词间的相似度都不高，且存在大量的无效数据；而 Amazon 与 Librarything 的重合度实验中，实验效果显著提高，这可能由于 Amazon 和 Librarything 都利用社会化标签而又是同语种。Amazon 和 Librarything 良好的表现，也促使我们试图做跨语言的重合度分析，一探究竟。当对标签进行分层比较后，有效数据和重合度的效果都有了较大的提高，提高标签质量或者增加用户的参与会增加标注系统之间的相似性并提高图书标注的准确性。不同类目下，英文的流行类、文学类的有效数据和重合度的表现都更好；中文的经管类和科技类则有更佳的表现。这些结果在不同语言用户的阅读习惯之间存在较大差异。

标签之间的重合度明显好于标签与主题词之间的重合度，而且提高标签频次会提高重合度效果。在不同语言之间，标签与主题词的重合度各有千秋，中文在经管和科技类的重合度更好，而英文在流行类和文学类的重合度更好。

4.2　结果启发

实验结果显示，社会化标签和主题词有其各自的优劣势，并在不同语言、不同类目之间又有不同之处。因此，我们试图将研究成果运用到实践中来。

社会化标签在揭示图书、用户参与等方面有其独特的优势，图书馆将社会化标签引入图书标注中无疑会增加用户对资源的关注。利用更加平民化的信息组织和检索方式，也会提升用户的满意度和图书馆的服务质量。同时，仍需要考虑社会化标签的非受控性。可以考虑将社会化标引和检索与传统的馆藏资源检索相结合，并行提供给用户，方便用户检索并提高检索效果。这样便能充分发挥传统资源组织形式和社会化标注系统二者的优势，提升图书馆服务，推进 Lib2.0 的发展进程。

5　结　语

Web2.0 的发展让我们看到了 Lib2.0 发展的必然趋势。社会标注系统作为新型的网络信息资源的组织方式越来越受到图书情报领域学者的关注和研究。社会化标签也开始进入图书馆工作领域，帮助提升用户体验。

本研究对中英几大社会化图书标注系统做了比较分析，发现了各大标注系统中社会化标签的异同。了解这方面的内容不仅能够帮助图书馆了解社会化标签，也能够为建立和完善书评网站提供借鉴。本文还引入了中国国家图书馆和美国国会图书馆的 MARC 主题词来和社会化标签进行对比。经过实验，发现社会化标签和主题词不管在个数、长度还是在相似度上都有较大的差异。图书馆若引入社会化标签，仍需要做大量的工作，以确保提升服务质量。

本课题还有很多的地方需要继续研究。社会化标签存在很多的近义词、多义词，仅仅语法层面的比较还不足以了解社会化标签的全部特征，特别是语义特征。因此，本课题还需要跨语言的标签分析，建立语义字典等手段强化分析，进一步挖掘社会化标签的特征和用户标签行为。

在对跨语言的比较中，我们仅比较了各同语种的标注系统。为了得到更加完整和可靠的比较分析结果，需要直接的跨语言的标注结果的比较。

参考文献：

[1] 宋彩萍，霍国庆. 信息组织论纲[J]. 中国图书馆学报，1997(1)：20－37.

[2] 徐少同，李书宁，徐文贤. OPAC 2.0 发展研究[J]. 图书馆论坛，2007，27(5)：86－88.

[3] DeZelar-Tiedman C. Exploring user-contributed metadata's potential to enhance access to literary works: Social tagging in academic library catalogs [J]. Library Resources and Technical Services, 2011, 55(4): 221－233.

[4] Tonkin E. Searching the long tail: Hidden structure in social tagging[C]//Proceedings of the 18th Workshop of the American Society for Information Science and Technology Special Interest Group in Classification Research. Austin: The American Society for Information Science and Technology, 2006: 1－10.

[5] Rae A, Siqurbjörnsson B, van Zwol R. Improving tag recommendation using social networks[C]//Proceedings of the 9th

International Conference on Adaptivity, Personalization and Fusion of Heterogeneous Information. Paris: Le Centre De Hautes Etudes InternationalesD'informatiqueDocumentaire, 2010: 92 - 99.

[6] Xu Zhichen, Fu Yun, Mao Jianchang, et al. Towards the semantic Web: Collaborative tag suggestions [C]//Proceedings of Collaborative Web Tagging Workshop at WWW2006. Edinburgh: University of Soathampton, 2006: 1 - 8.

[7] Lee S O K, Chun A H W. A Web 2.0 tag recommendation algorithm using hybrid ANN semantic structures [J]. International Journal of Computers, 2007(1): 49 - 58.

[8] 钟青燕,苏一丹,梁胜勇. 基于层次聚类和语义的标签推荐研究[J]. 微计算机信息,2010(3 - 6): 199 - 203.

[9] 刘知远. 基于文档主题结构的关键词抽取方法研究[D]. 北京:清华大学,2011.

[10] Brooks C H, Montanez N. Improved annotation of the blogosphere via autotagging and hierarchical clustering[C]//Proceedings of the 15th International Conference on World Wide Web. Edinburge: ACM, 2006: 625 - 632.

[11] Shilad S, Jesse V, John R. Learning to recognize valuable tags [C]//Proceedings of 14th International Conference on Intelligent User Interfaces. Sahibel Island: ACM, 2009: 87 - 96.

[12] Rolla P J. User tags versus subject headings: Can user-supplied data improve subject access to library collections [J]. LRTS, 2008, 53(3): 174 - 184.

[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] Lu Caimei, Park J, Hu Xiaohua. User tags versus expert - assigned subject terms: A comparison of LibraryThing tags and Library of Congress subject headings [J]. Journal of Information Science, 2010, 36 (6): 763 - 779.

[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.

[16] Wu Dan, He Daqing, Qiu Jin, et al. Comparing social tags with subject headings on annotating books: A study comparing the information science domain in English and Chinese[J]. Journal of Information Science, 2013, 39(2): 169 - 187.

Comparative Study on Books Social Tags of Chinese-English Translation

Lu Chao Zhang Chengzhi

Department of Information Management, Nanjing University of Science and Technology, Nanjing 210094

[**Abstract**] With the development of Web2.0, social annotation system has affected the traditional information organization in libraries. Taking the books of Chinese-English translation as the research objects, this paper collects social tags from Douban, Amazon and Librarything, and gets the subject headings of MARC records from National Library of China and Library of Congress, to explore the differences between tags and subject headings of two types of vocabulary of each book in the length, number and the similarity. The results show the differences in length, number and similarity of social tags or subject headings of different languages, differences in those of books in different catalogues, and differences between tags and subject headings of a single book. This study can help the libraries to have a good knowledge of social tags and improve the service with them.

[**Keywords**] social tag book annotation subject heading social annotation

下 期 要 目