Social Tagging in Cross-Culture and Cross-Language Environments: A Comparative Study

跨文化多语言环境下大众标注比较研究

徐晨◎著

湖南科学技术出版社

本书的研究和出版获以下研究项目的资助：

国家自然科学基金青年项目（71103203）：
多语言环境下Social Tagging的内涵机理与应用框架研究；

2011年度中央高校基本科研业务费青年教师助推专项（2011QNZT237）：
多语言环境下Social Tagging应用模式研究

跨文化多语言环境下大众标注比较研究

徐晨◎著

CTS 湖南科学技术出版社

图书在版编目（CIP）数据

跨文化多语言环境下大众标注比较研究 / 徐晨著.
-- 长沙 : 湖南科学技术出版社, 2012.10
ISBN 978-7-5357-7434-7
Ⅰ. ①跨… Ⅱ. ①徐… Ⅲ. ①社会习惯语－网络检索－对比研究 Ⅳ. ①H034②G354.4
中国版本图书馆 CIP 数据核字(2012)第 233628 号

跨文化多语言环境下大众标注比较研究

著　　者：徐　晨
责任编辑：汤伟武
出版发行：湖南科学技术出版社
社　　址：长沙市湘雅路 276 号
　　http://www.hnstp.com
印　　刷：深圳市深博数码科技有限公司
　　（印装质量问题请直接与本厂联系）
厂　　址：深圳市福田区八卦三路 429 栋二楼西 203
邮　　编：518029
出版日期：2012 年 10 月第 1 版第 1 次
开　　本：850mm×1168mm　1/32
印　　张：6.125
字　　数：165700
书　　号：ISBN 978-7-5357-7434-7
定　　价：28.00 元

序

大众标注真正兴起是从2003年底Delicious（原用名是Del.icio. us）提出社会书签（social bookmarking）的概念开始的。运用个性化标注方式，鼓励用户通过标注词（Tags）保存、组织、查找和管理自己的网络资源，这种曾在20世纪末出现萌芽的运用才重现于因特网，并因其与网络其他应用如博客（Blog）、维基百科（Wiki）等的相似特征被一并归入Web 2.0应用。

如今，大众标注已进入广泛使用、迅速发展时期，但标注词的语言特征和标注行为的社会性，使得大众标注比其他Web 2.0应用方式更多地涉及文化和社会因素。不过纵观大众标注整体研究状况，将其置于不同文化和社会背景下进行研究还是一个较大的空白。

徐晨在武汉大学攻读博士学位期间，有机会于2007～2009年到美国学习研究两年，她敏锐地发现不同语言文化对大众标注的影响，选择具有代表性的中美标注网站及其站点所抽取的标注词作为嵌入口，研究探讨不同文化社会背景下的大众标注，整合中美大众标注在功能设计、使用特点、词语组织这三个比较视角的研究成果，提出如何结合不同的组织语言特点，兼顾不同文化思维方式和不同语言特征，发展跨文化跨语言背景下大众标注的一系列有针对性的建议和措施，可以说在大众标注研究领域进行了富有新意的尝试和探索。

本书具有以下特色：

第一，选择Delicious和365Key作为对象，从标注辅助属性、标注显示方式、标注的相互链接、标注使用和管理四个角度全面比较总

结中美大众标注站点在功能设计上的区别和联系。根据比较的分析结果，总结了 Delicious 优于 365Key 的三个方面，并发现适合的设计功能可更好地服务大众标注站点用户创制使用和共享重用标注词。

第二，使用网络编程工具 Python 抽取 Delicious 和 365Key 一定时段新闻类的标注词，从使用规律、使用问题、词性、词型、词义和类别六个不同的使用进行综合比较，揭示出这些方面两个站点的各自特色，并从中挖掘出更深层次中英文语言特点和东西方文化思维方式的差异。

第三，选择 Library of Congress Subject Headings（LCSH）和《中国分类主题词表》（《中分表》）作为中美受控语言的代表，分别将已抽取的 Delicious 和 365Key 的标注词与来自 LCSH 和《中分表》的受控词进行了各自匹配比较。结果发现标注词与受控词差别较大，主要受不同组织语言特点影响，同时与词表、功能设计、语言特征等多种差异相关。

第四，整合中美大众标注在功能设计、使用特点、词语组织这三个比较视角的研究成果，提出如何结合不同的组织语言特点，兼顾不同文化思维方式和不同语言特征，发展跨文化跨语言背景下大众标注的一系列有针对性的建议和措施。

徐晨在武汉大学学习期间，我曾是她的硕士和博士指导教师，我们在一起学习交流六年。她勤奋刻苦、认真踏实、思维敏锐，具有探索精神，取得了很好的成果。

随着微博、社会媒体和各类网络社会化应用的兴起，大众标注又迎来了一个新的发展机会。期盼作者能在已有研究成果基础上，将视角从文本引入更广泛的社会媒体类型，探索大众标注在社交网站中的新发展形式和应用模式。

2012 年 7 月于武昌珞珈山

目 录

第1章 绪 论

1.1 研究背景

大众标注（social tagging）是运用先行于概念，它的前身是共享性的网络书签（online bookmarking），从1996 it List把私有书签正式作为一种功能提供给用户到1999年Backflip，Blink，Clip2等提供网上文件夹进行书签管理（有些站点甚至可以自动存储书签到文件夹中），网络书签经历了竞争发展的三年，不过21世纪初网络泡沫的破灭，使得这一很有潜力的运用消失于萌芽中。直到2003年年底Del. icio. us[①]提出社会书签（social bookmarking）的概念，创造性地运用个性化标注方式，鼓励用户通过标注词（Tags）保存、组织、查找和管理自己的网络资源，这种运用才重现于因特网。2004年随着Del. icio. us在网上的火爆，很多大众标注网站陆续兴起，Flur，Citeulike，Connotea从不同方面强化了大众标注的使用深度，扩展了大众标注的功能与范围。一些信息专家看到这种运用的巨大潜力，于是将它与一些呈现类似特征的网络运用如Blog，RSS等一同归入Web 2.0的范畴。

从发展历程看，大众标注最初只是被当做一种在线的网络收藏夹，是一些技术爱好者为网络用户提供除IE收藏夹以外另一种存储再用网络资源的便利手段。不过随着网民开始创造一些个性化的

①Del. icio. us自2005年被Yahoo收购后，其网名改为Delicious. com

书签（bookmarking）来标注他们的资源，这一标注方式的灵活自由性和简单大众化逐渐吸引了越来越多的用户加入标注者（taggers）的行列，加上它与社会网络的天然结合性，大众标注作为 Web 2.0 一种重要应用方式已成为互联网用户广泛的使用行为。根据 2007 年美国知名统计机构 Pew Internet & American Life Project 的专家 Rainie 所作的统计报告，已有 28% 的在线美国用户使用大众标注[1]。同时在大洋彼岸的中国，大众标注站点也开始吸引越来越多的网络用户。上网进行网络资源的标注已成为一种社会现象，普遍存在于网络世界中。

伴随着大众标注用户面的普及扩大，一批反映用户不同文化语言特点并来源于日常生活的习惯用语已形成一种聚集趋势，被广泛用来描述、组织和共享使用各类网络信息资源。大众标注已不单纯是一种标注方法，更成为人们使用和管理网络资源的重要方法，并且这些词语特点映射出不同语言文化背景下用户的不同使用特点和组织方式。2004 年底，Thomas Vander Wal 在他的博客文章中提出 Folksonomy 这一概念，他把“Folk”和“Taxonomy”联系起来，创造了 Folksonomy 这个词语[2]。大众分类法（Folksonomy）作为一种新型的信息组织方式由此正式诞生。不同于以往由专家或专业人士发布相关词表这种从上至下的信息组织方式，信息组织方式是由民众自主创造标注词汇而逐步聚合形成一批常用的标注词，来描述不同类型的信息。不少研究人员开始逐步探讨这种新的组织方式的机制和使用方式，宏观上比较它与传统组织方式的异同[3-5]。但从本研究目前所收集的文献资料来看，把这两种组织方式放在不同文化背景下进行纵向的比较研究还很缺乏，尤其是具体将传统受控词汇与大众标注词汇进行语义、语法和语用的多维比较在美国还属于尚待研究的问题，在中国更是鲜有这方面的研究出现。

1.2 研究历程和现状

随着大众标注和大众分类法等概念的产生，一批学者们开始在博客和期刊中发表一些介绍性文章，推广和普及这种极具潜力的应用和信息组织方式。Darlene Fichter 是早期大众标注的先驱作家，她发表过一系列有关 Web 2.0 和大众标注的文章。她 2004 年就介绍过 Delicious 等工具的使用情况，展望了大众标注蓬勃发展的前景[6]，2006 年又断言大众标注将成为现代人们网络生活一部分，并发展成为一种普遍的社会现象[7]。2004 年 12 月，Adam Mathes 和 David N. Sturtz 分别发表了推广大众标注的重要论文。Adam Mathes 分析大众分类法优势之处在于它不同于以往专业人员和作者创建的元数据，并以其在 Delicious 和 Flicker 中的运用情况为例，探讨了这种新型用户创造性的分类方式产生原因、工作原理和未来的发展趋势[8]。David N. Sturtz 分析了大众分类法的内涵、当前应用情况和未来研究方向[9]。

2004 年大众标注概念开始得到广泛的认识，到了 2005 年大众标注的用户群更是迅速扩大，更多的介绍性文章和综述性文章开始在期刊上发表。The Educause Learning Initiative's (ELI's) 发表了 "7 things you should know about social bookmarking"，这篇文章从七个方面概括了社会书签 (social bookmarking)，首先虚拟了一个场景介绍社会书签的使用情况，接着从含义、使用者、工作原理、使用意义、运用趋势、应用范围以及对教育和学习的影响七个方面进行了全面介绍，此文虽然短小，但简明扼要，概括也较为全面，是一篇不错的描述社会书签入门级文章[10]。Hammond 等在 *D-Lib Magazine* 发表两篇针对大众标注工具较全面的综述性文章，第一篇文章围绕着大众标注的产生背景和发展历程，概括了它的定义和特征，分析了它与传统分类的关系，并引申剖析了这种迅速发

展的网络应用能吸引用户使用的心理原因以及目前应用中的优点和不足，附录中还对当时已有的大众标注站点进行了总结和概括[11]。其后第二篇文章，Lund，Hammond 等又以 Connotea 为例，实证分析了大众标注系统的架构和使用原理，通过具体例子展望了大众标注的改进措施和发展情况[12]。Emma Tonkin 介绍了大众分类法的产生和兴起，联系元数据分析了大众标注的特征和潜力[13]。Laura Gordon-Murnane 比较大众标注相关网站的性能和各方面特征，讨论了工具发展的目标[14]。Jessica Dye 对 David Weinberger、Peter Morville 等知名专家的博客文章进行了总结评述，也对 Yahoo 和 Amazon 运用大众标注情况进行了归纳分析[15]。Amanda Etches-Johnson 介绍了大众标注的社会性特征，展望了大众标注与 OPAC 结合的可能[16]。Edith Speller 尝试着对大众标注作了一个从概念到过程再到应用的文献评述，可惜篇幅较短，涵盖面有限[17]。同时期刊杂志也涌现出一批推广 social bookmarking 使用的介绍性文章，如 Mary Ellen Bates 就在一篇简短的文章中具体描述了 Flur 的使用步骤[18]。

随着大众标注的影响力日益增加，不少国际会议也开始把大众标注作为它们的会议主题进行探讨。2005 年 10 月，在纽约召开的由英国 UKLON（The UK Office for Library and Information Networking）举办的“第六届网络信息系统工程国际大会”，其中有些研究专题涉及大众标注[19]。2006 年，JCDL（Joint Conference of Digital Libraries）和 ASIS&T（The American Society of Information Science and Technology）年会也有相关论文发表，但比较局限于发展情况和内在规律的探讨[20-21]。2007 年，在 Milwaukee 召开的 ASIS&T 年会的主题为“合作与研究实践：社会计算与信息科学”，其中探讨了不少大众标注的主题，有讨论大众标注的行为和运用范围[22]，有专论它与信息检索的关系[23]，有挖掘它在合作工作中的影响和用途[24]，有论述图像站点 Flicker 的标注词的使用

规律[25]。

目前大众标注研究逐步深入，已从早期围绕着发展情况、概念及描述向注重应用、模型探讨和领域结合的态势发展。专家们普遍意识到大众标注是一种未来信息资源组织和管理有用的方式和工具。于是信息科学、社会科学、心理学、计算机科学等多个领域的专家开始把大众标注融入他们熟悉的专业范畴中进行研究，力求利用大众标注为本领域的理论和实践服务。例如，信息科学专家主要探讨大众标注如何嵌入当前信息组织方式，包括大众标注与本体语义网的联系，信息检索如何运用大众标注，大众分类法与传统分类法的关系等，在后续的相关研究评述中将对它们展开具体评述。社会科学专家和心理学学者们主要关注大众标注的用户使用情况，用人机交互理论解释大众标注的用户方面问题[26-30]。虽然用户研究也是大众标注一个关键的研究方向，但这个领域范围很广，笔者时间和精力有限，不可能面面俱到，因此本论文仅就标注词的特征分析用户的使用特点。

专家们对大众标注的态度从最初的推崇到批评再到如今理性看待它，既已认同它自由灵活的优势，也意识到其不规范散乱的劣势，研究思路集中于如何努力挖掘大众标注的内在潜力，借鉴传统组织方式的优势，促进大众标注的整体发展，使其为不同文化背景用户提供信息服务。

1.3 研究目的和意义

大众标注如今已进入广泛使用、迅速发展时期，但标注词的语言特征和标注行为的社会性，使得大众标注比其他 Web 2.0 应用方式更多地涉及文化和社会因素。不过纵观大众标注整体研究状况，将其置于不同文化和社会背景下进行研究还是一个较大的空白。

不少研究人员开始逐步探讨这种新的标注方式的机制和使用，

宏观上比较它与标引（indexing）的异同。其实大众分类法出现是为传统分类方法注入了一种活力，其存在的必要性是不可置疑的，当然不足之处也是不可忽视的，关键是如何为这两种组织方法找到一个很好的结合点，让它们互相补充，扬长避短。

但是整体而言，把这种组织方式放在不同文化背景下进行纵向的比较研究还很缺乏，尤其是具体将传统主题词表与大众标注词汇进行语义、语法和语用的多维比较在美国还属于尚待研究的问题，在中国更是鲜有这方面的研究出现。

针对目前大众标注的研究现状，文章将以具有代表性的中美标注网站（如 Delicious 和 365Key）及其站点所抽取的标注词作为嵌入口，研究探讨不同文化社会背景下的大众标注。大众标注虽起源于美国，但 Delicious 等美国网站用户非常广泛，并不局限于美国本土范围，而是遍及世界各地。因此论文标题以跨文化和跨语言更准确地概括比较的维度和环境。这里的“跨语言”不仅指中文和英文两种不同网站通用语言，还指标注词和受控词这两种信息组织语言。跨文化则是将以中英文所代表的典型东西方文化作为比较对象。不过为了研究方便，兼之选取的网站来源于中美两国，本论文中将把“中美”和“跨文化、跨语言”两个词组视为同义，并交替使用。毋庸置疑，美国大众标注的整体发展速度要快于中国，通过中美站点功能和界面设计比较研究，可以相互借鉴，促进大众标注系统的整体发展创新。同时通过比较中美大众标注站点的标注词语，发现这种标注语言背后折射出的文化差异。然后将标注词与受控词汇［即《美国国会主题词表》LCSH（Library of Congress Subject Headings）和《中国分类主题词表》］中的主题词进行匹配比较，由此探讨大众标注与传统的组织方式的各自特点、异同及其产生原因。最后整合中美大众标注在功能设计、使用特点、词语组织这三个比较视角的研究成果，提出如何结合不同的组织语言特点，兼顾不同文化思维方式和不同语言特征，发展跨文化跨语言背

景下大众标注的一系列有针对性的建议和措施。

1.4 研究问题的提出

根据上述研究目的，本研究将探讨以下问题：

- 问题1：中美大众标注网站在设计、功能等方面有何异同？
- 问题2：中美大众标注在词语类型、内容和使用上有何异同？
- 问题3：中美大众标注词与来自传统主题词表的主题词有何异同？大众标注和传统的主题词标引（indexing）在多大程度上可以互补？
- 问题4：通过比较研究可为整个大众标注发展提供什么建议？中国和美国的大众标注之间有何可互相参考借鉴之处？

1.5 全书结构编排

第1章“绪论”。本章主要描述大众标注发展背景，展现大众标注的研究现状和不足，从而阐述文章研究目的，进而重点提出研究的问题，说明文章的整体结构安排。

第2章“大众标注相关研究综述”。本章会先对国内外关于大众标注研究的所有科研文献进行汇总统计，按年度展现了国内外研究趋势。然后主要针对研究问题从大众标注网站分析、标注词的分析研究和大众标注与传统组织方式的比较三个方面的相关领域及研究分支即大众标注网站比较研究、大众标注系统用户使用研究、标注词的使用模型、标注词的类型、大众标注语言文化特征研究、大众标注与传统信息组织方式的比较等方面的相关理论论述和研究成果进行总结和评述，指出这些方面研究的可借鉴之点及其不足之处，为全文的研究提供理论依据和方法论的基础。

第3章“跨文化多语言背景下大众标注比较研究的相关问题和

范畴”。基于第 2 章的文献评述，针对当前研究不足，进一步阐述本论文的研究具体目的，详细阐述第 1 章提出的研究问题研究要点和具体内容，并解释相关研究变量在本研究中的具体含义。

第 4 章“跨文化多语言背景下大众标注比较研究的方法和步骤”。本章针对第 3 章提出的研究问题，在研究范畴界定的基础上，具体逐一介绍选取的各种研究方法及其在本研究中的使用特点，同时说明研究数据来源和选取原因。在确定了研究方法和数据来源基础上，详细描述整个研究设计的数据收集过程，并集中阐述将如何对采集而来的数据进行分析。

第 5 章“跨文化多语言背景下大众标注比较详析”。基于笔者发表于 2008 年 ASIS&T 年度会议录中所做的前期测试研究，基于前几章提出的研究问题和研究方向，进行三个方面的分析：首先对选取的中美大众标注站点的网站功能设计进行了详细比较，其次对两个站点所抽取的标注词进行横向对比，最后将这些标注词与其所在国家的受控词表进行纵向的比较分析。

第 6 章“跨文化多语言背景下大众标注比较的结论与展望”。根据第 5 章的数据分析和讨论，总结三个方面所取得的研究成果，提出如何结合不同的组织语言特点，兼顾不同文化思维方式和不同语言特征，发展跨文化跨语言背景下大众标注的一系列有针对性的建议和措施。同时说明本论文的研究局限所在，并展望未来进一步研究的方向和思路。

第 2 章　大众标注相关研究综述

从论文发表时间上看，国外论文发展趋势基本呈现一种快速发展的曲线，2004 年的研究还不够成熟，所以只有若干文章发表在一些知名学者的博客中，但这毕竟是一种开创性的尝试，所以后来对大众标注的概念阐述基本延续了当时的思路。2005 年陆续有部分相关论文发表在相关期刊中，不过这个时候专家们对大众标注的未来发展还未有一个明确的方向。2006 年是稳步发展阶段，专家们既看到大众标注在各个信息领域方面的巨大潜力，又对其存在的问题颇为头疼。2007 年和 2008 年是大众标注的飞速发展时期，相当一批有深度的文章涌现，主题深入到信息管理和服务的各个方面。当然这一趋势也可以通过相关数据进行佐证。笔者对 H. W. Wilson、LISA、LISTA、ISI、ACM 几个数据库中刊发于同行评议期刊上的相关英文文章进行了书目统计，2005 年仅有 5 篇相关文章，2006 年迅速增长到 29 篇，2007 年快速增长到 59 篇，2008 年稳步增长到 76 篇。

同样从时间维度上看，国内对大众标注的研究数量也一直呈现迅速增长趋势，从 2005 年的仅有 2 篇到 2006 年慢慢增长到 7 篇，然后 2007 年以两倍于前的速度增至 16 篇，再到 2008 年的 26 篇，增长幅度基本呈现一种直线飙升的趋势。在数量增长的同时，短短几年时间，国内研究深度也有了较大提高。严格地说，2005 年时的论文还算不上真正意义的研究论文，充其量只是粗略对大众标注的介绍和说明。步入 2006 年，国内图书情报专家学者才真正着眼于大众标注的研究。这批文章虽仍以介绍大众标注为主，但已从理