



湖北经济学院学术文库

社会标签推荐技术及其应用研究

SHEHUI BIAOQIAN TUIJIAN JISHU JIQI YINGYONG YANJIU

■ 靳延安 著



湖北经济学院学术文

社会标签推荐技术及其应用研究

SHEHUI BIAOQIAN TUIJIAN JISHU JIQI YINGYONG YANJIU

■ 靳延安 著

图书在版编目(CIP)数据

社会标签推荐技术及其应用研究/靳延安著.
武汉:湖北人民出版社,2016.6
ISBN 978-7-216-08957-9
I. 社… II. 靳… III. 算法设计 IV. TP301.6
中国版本图书馆 CIP 数据核字(2016)第 157699 号

责任部门:高等教育分社
责任编辑:陈晓东
封面设计:张 弦
责任校对:范承勇
责任印制:王铁兵

出版发行:湖北人民出版社	地址:武汉市雄楚大道 268 号
印刷:武汉市福成启铭彩色印刷包装有限公司	邮编:430070
开本:710 毫米×1010 毫米 1/16	印张:8.5
字数:150 千字	插页:2
版次:2016 年 9 月第 1 版	印次:2016 年 9 月第 1 次印刷
书号:ISBN 978-7-216-08957-9	定价:25.00 元

本社网址:<http://www.hbpp.com.cn>
本社旗舰店:<http://hbrmcbs.tmall.com>
读者服务部电话:027-87679656
投诉举报电话:027-87679757
(图书如出现印装质量问题,由本社负责调换)

总 序

1919年，当中国这个东方文明古国正在经历五四新文化运动洗礼时，德国著名社会学家马克斯·韦伯在慕尼黑大学进行了一次题为《以学术为志业》的演讲。在百年的曲折前行中，韦伯的演讲像暗夜中的一盏明灯，给无数中国学者带来希望之光，影响深远。

面对当下社会转型、信仰缺失等矛盾与冲突，每一位有志于学术的学者都需要重新审视自己：“我们为什么要从事学术？”法国哲学家米歇尔·福柯说：“有人会悲悼目前的真空，在理念的世界中继续追寻一种专制。但是另外一些人，一旦在他们的生命中发现了一种新的音调，一种新的观看的方式，一种新的行为方式，我相信这些人决不会哀叹这个世界是一个谬误，不会哀叹历史上充满了无足轻重的人，不会喝令别人住嘴以使自己的责难之声被人听见。”福柯的一番话解答了无数学者心中的困惑，指出了学术研究绝不仅仅是一种职业，也是一种生活方式，甚至是一种游离于世俗价值之外的信仰，一种在繁华落尽之后仍能找到生命意义与希望的精神自足与崇高。

以学术为志业，要将生命融入学术，淡泊名利。国学大师梁启超在《清代学术概论》一书中将清代学者取得巨大成就的主要原因归结为：“俗既俭朴，事畜易周，而寒士素惯淡泊，故得与世无竞，而终其身于学。”一个人只有超越名利的束缚，才能洞察宇宙人生的奥妙。韦伯曾言“学术生涯是一场鲁莽的赌博”，但若每一位有志于学术的人具有“昨夜西风凋碧树，独上高楼，望尽天涯路”的心理准备与承受孤独的勇气，同时有“衣带渐宽终不悔，为伊消得人憔悴”的坚持，那么他一定会体验到“众里寻她千百度，蓦然回首，那人却在灯火阑珊处”的喜悦与满足。孔子云：“朝闻道，夕死可矣！”更是将生命与学问相联系。事实上，如果学术不能上升到信仰的高度，不能融入生命意义与人生价值的话，要么变得空洞无物，要么成为沽名钓誉的工具。

以学术为志业，要怀经世济民之心，心系天下。很多人感慨：今天的社会很难再产生大师！是我们的物质条件太差？还是自身对学术失去了敬畏之

心？或许，我们可以从追溯往哲先贤的足迹中寻找答案。中国古代士人素有关怀天下的传统，宋代理学家张载将学术的追求归纳为四句话：“为天地立心，为生民立命，为往圣继绝学，为万世开太平。”这是何等的气魄和胸襟！在民国那个曾经大师辈出的年代，鲁迅、闻一多……一个个响亮的名字背后，是浓浓的战火，弥漫的硝烟。正是将思想与学问着眼于整个民族的前途命运、生死存亡，大师们才取得了傲人的成绩。心系国家，才会以神圣而虔诚的态度面对学术，才会产生经世济民的不朽著作。反之，一心只为“稻粱谋”，纵然有万般才智，也很难成气候。

“大学者，研究高深学问者也”，学术永远是大学的灵魂。作为一所合并成立仅十余年的年轻高校，湖北经济学院可能刚刚入门，谈不上鲜明的学术传统，但我们有巨大的发展潜能。建校至今，我们始终保持着向上的活力与动力。一大批以学术为志的年轻博士、中青年学者踊跃加盟，推动学校科研实力持续提升，高水平研究成果不断涌现。

一套丛书不仅可以展现一系列的学术成果，还能展现一所学校、一个国家，甚至一个时代的学术与思想风貌。正如我们观《永乐大典》，可以感受明代盛世的威仪与气魄；览《四库全书》，可以了解清代学术的整体格局。我们希望通过“湖北经济学院学术文库”向社会展示学校的学术风貌与氛围。因此，我们遴选著作时，首先坚持精品原则，努力推出高水准的学术著作，宁缺毋滥。其次，我们坚持百家争鸣，百花齐放，鼓励不同学科、不同学术风格的著作登上舞台。最后，我们希望一部著作不仅表达作者的学术观点，还能体现各自的学养与治学态度。严谨务实，充满个性的文章除了会给读者学术上的启迪之外，还能让人洞察到文章背后鲜活的人格与个性魅力。

学术文库建设没有终点，每一位以学术为志业的经院人都可以进入这个行列。薪火相传，愿“湖北经济学院学术文库”能成为一道亮丽的风景线，并以此作为我们新的舞台和新的学术起点，伴随着湖北经济学院的成长与腾飞，愈加绚丽多姿！

湖北经济学院院长：漆腊应

目 录

1 绪论	1
1.1 研究背景.....	1
1.2 社会标签推荐系统主要研究内容.....	3
1.3 本书研究的主要内容及成果.....	6
1.4 本书的组织结构.....	7
2 标签及社会标签系统.....	9
2.1 标签和社会标签.....	9
2.2 社会标签系统的研究对象	13
2.3 社会标签的标注方法	17
2.4 社会标签的浏览和可视化	18
2.5 社会标签与信息资源的组织及利用	19
2.6 社会标签系统实例	26
2.7 本章小结	27
3 推荐技术及相关理论	29
3.1 传统推荐技术	29
3.2 推荐效果评价方法和评价指标	47
3.3 相似度计算	55
3.4 社会标签推荐系统面临的问题	56
3.5 本章小结	62
4 基于不同粒度的标签推荐	63
4.1 引言	63
4.2 基于词粒度的标签推荐	64
4.3 基于话题粒度的标签推荐	68

4.4 基于混合粒度的标签推荐	73
4.5 实验分析	75
4.6 本章小结	80
5 基于话题敏感的标签排序	82
5.1 引言	82
5.2 基于话题敏感的标签排序	83
5.3 实验与分析	89
5.4 基于话题敏感的标签推荐	92
5.5 本章小结	95
6 基于用户动机倾向性的标签推荐	96
6.1 引言	96
6.2 用户标注动机倾向性的类型及度量	97
6.3 用户标注动机倾向性判别	103
6.4 基于用户动机倾向性的标签推荐	105
6.5 实验与分析	110
6.6 本章小结	113
7 总结与展望	114
7.1 主要贡献	114
7.2 工作展望	115
参考文献	117
后 记	129

1 绪论

1.1 研究背景

随着 Internet 的日益发展,网络信息资源正以人们难以想象的速度增长,Web2.0 的出现使得这种增长势头更加迅猛。在 Web 2.0 中,互联网体系由来自上而下的、少数资源控制者集中控制、主导转变为自下而上的由广大用户集体智慧和力量主导。用户除了是网络信息资源的浏览者之外,同时还是网络信息资源的生产者。Web2.0 用户创造内容的这种特性虽然丰富了信息的来源,加速了信息的扩散,但也同时引发了信息过载、搜寻负荷加重、信息品质降低等问题。那么,用户作为生产者如何方便地、容易地为内容贴上合适的标签,作为管理者如何有效地、合理地组织和管理网络信息资源,作为消费者如何快速地、低成本地、获取合适而又高质量的信息就成为了摆在我们前面的不可逾越的重大研究课题。

理想的网络信息资源组织应该以用户为中心,充分利用新兴技术和人们积累的经验,组织体系应具备较高的实用性和易用性。在 Web2.0 环境下,大众分类^①作为网络信息资源组织的一种非常有效的方法正发挥着重要的作用。作为组织方式,它与传统的自上向下的、刚性的受控层次分类体系不同,大众分类体系具有以下三个优点:(1)标签^②是网络信息资源用户对网络资源标注时产生的,相同的标签经过汇集以后形成了新的分类,它是自下向上的;(2)标签不受控于专家,用户可以自行使用任意词进行标注,具有极高的灵活性、易用性和

① 大众分类(Folksonomy)是由“Folks”和“Taxonomy”组合而来的,“Folksonomy”是指“群众”自发性定义的平面型非层级结构式标签分类,也指使用者可以用任意关键字进行分类的协同工作。现有“分众分类”“通俗分类”“大众分类”等不同的翻译名称。

② 本文中,为了区分用户(或者称为标注者,tagger)的标注行为(tagging)和标注结果(tags),将用户进行标注的行为称为“标注(social tagging)”,所选的要赋予网络资源的标注结果成为“标签(social tags)”。

主观认知性,网络资源可以“柔性”隶属于多个大众分类。(3)在大众分类体系中,用户可以从多个维度、多个层次对网络资源进行标注。因此,它的结构是非层次的。

然而在具备众多优势的同时,标签方式同样存在缺点,主要表现于以下两个方面:(1)大多数社会标签系统允许用户自行输入标签,这种运作模式使用户很容易控制标注行为,但由于标注的随意性也造成了标签中存在较多噪音,错拼、歧义以及无实际意义的用户自定义标签常常充斥其中,这对标签的实用性造成了不小障碍。为此,有些社会标签系统不得不专门为用户给出一些指导原则^[1]。(2)数据稀疏问题,由于标签式浏览是一种新兴的信息组织方式,还没有得到十分广泛的应用,尤其在中文资源中,采用这种组织方式的网络资源十分稀少,另一方面,用户尚不习惯于为网络资源添加大量的标签,从而使得网络上现有的标签资源极为稀少。

近年来,正是在这种实际需求下,标签推荐技术受到了学术界以及互联网企业的广泛关注。如果以“tag recommendation”“social tagging”为关键字进行精确查找,在 Google 学术搜索上查找到 2005 年以来的文章 19400 篇次。国外关于社会化标注系统的研究开始于 1999 年,在 2006 年以前稳步增长,之后进入快速发展时期,到了 2009 年,达到该领域的文献量高峰期,共 138 篇,然后又有所下降,2011 年的文献量为 59 篇,截至目前,文献数量呈稳步上升趋势,说明在 2009 年以前该领域经过了快速成长期,然后进入成熟期,由于对社会化标注系统存在的标签同义、多义等现象,所以如何改进标签系统成为这一时期的瓶颈问题,2010 年以后,学者们主要利用各种工具对标签的语义加以正确分析,进展比较缓慢。可见,标签推荐在学术界已经有了一定的研究热度。同样,在企业界也出现了各种各样的社会标签系统,比如 Delicious、Flickr、Bibisonomy、豆瓣等。

标签推荐就是通过考察、分析、挖掘网络信息资源的内容和用户的标注历史、显式或隐式的关系为待标注网络信息资源提供一系列高质量的标签作为候选。这些候选标签按照相关程度由高到低降序排列,推荐时选择前面的标签进行推荐。推荐的目的是:(1)简化标注程序,方便用户使用,从而增加社会标签系统的可用性和黏性。(2)提高标签的质量,减少错拼、歧义等情况,提高标签在信息资源组织、检索、利用和发现的作用。(3)改变标签空间的结构,使得标签空间更快的稳定和收敛,进而涌现语义。

综上所述,社会标签系统在理论和实践中都得到了很大的发展,但随着互联网应用的发展,社会标签系统也面临着一系列挑战。针对社会标签系统面

临的主要挑战,本书将对社会标签系统中的若干关键技术进行有益的探索和研究。

1.2 社会标签推荐系统主要研究内容

推荐系统的相关技术研究可以追溯到其他相关领域的广泛工作,例如认知科学、近似理论、信息检索、预测理论,甚至与管理科学也有联系。随着互联网的普及和电子商务的发展,推荐系统逐步成为电子商务中一项重要的研究内容。特别是自第一篇关于协作过滤^[2]的论文发表以来,推荐技术研究就备受学术界的关注。ACM 于 1999 年设立的电子商务兴趣小组 SIGECOM 就开始关注推荐技术,在第一次会议上有关电子商务商品推荐的文章占了很大的比重。同样隶属于 ACM 的 SIGKDD 兴趣小组于 1999 年开始设立 WEBKDD 研讨组,研讨主题为电子商务系统中的推荐技术。1999 年召开的人机交互会议 (CHI1999) 专设 Workshop 讨论推荐系统中的交互。作为电子商务领域的重要技术,许多学者提出了至今都在广泛应用的推荐算法。后来相继有 IJCAI、SIG-IR、WI、SIGKDD、IUI、AAAI、EC-WEB 等著名的国际学术会议设立专题讨论推荐系统。从 2006 年夏天开始,ACM 开始设立专门研究推荐系统相关的会议 RecSys,至今已经召开六届。

推荐技术研究之所以保持高涨的态势,一方面主要因为它是面向问题领域的。现实生活中有许多领域都需要推荐,比如在亚马逊网站(Amazon.com,如图 1-1)上推荐书、光碟和其他产品,在豆瓣(Douban.com,如图 1-2)上推荐电影,在美味网(Delicious.com,如图 1-3)推荐标签等。另一方面,因为有许多需要处理“信息过载”和“个性化服务”的需求。尽管搜索技术大行其道,可以让用户主动地从“信息过载”的负担中解放出来,找出自己需要的、目标较为清晰的信息,但是让用户的互联网体验更加个性化,帮助用户轻松“发现”他感兴趣的信息,或者说“让信息找到他”,就必须认识到“搜索已死,推荐上位”^[3]。推荐功能已经逐渐成为面向用户的互联网应用的标准配置,或将可能会发展成为互联网应用的基础设施。

这本书的其他版本 ····· (全部14)

Penguin Books Ltd版 2006 / 489人读过

湖南科学技术出版社版 2008-9 / 3189人读过 / 有售

湖南科技版 2006-11 / 952人读过 / 有售

Farrar, Straus and Giroux版 05 April, 2005 / 450人读过

以下豆列推荐 ····· (全部)

黄集伟的一架好书 (饥饿艺术家)

互联网产品经理 全方位入门 (iamsujie)

罗振宇《罗辑思维》推荐书籍 (怪刺)

曾改编成电影的书 (信而好古)

豆瓣上最热的财经图书 (鱼鱼)

谁读这本书?



囡囡爆发小宇宙
27分钟前 想读

图 1—1 豆瓣网图书推荐示例

Bet the FIFA World Cup | New and odds for the 2014 FIFA World Cup in Brazil

Title

Brazil 4-star, football, sportsbetting, worldcup,

Tags ?

Separated by comma

football soccer sportsbetting worldcup

Suggested Tags ?

Comment

1000

Public

Twitter

Facebook

Cancel

Save link

图 1-2 美味网标签推荐

社会标签推荐系统与传统的推荐系统相比,主要有两点不同。以电子商务推荐系统为例来说明它们的不同。首先,最大不同点在于推荐对象的不同。社会标签系统中的推荐对象是标签,而传统的电子商务系统的推荐对象是具体的商品。标签本身具有语义特性,可以反映出用户的偏好,比如一个人对电影《泰坦尼克》的标注可能是“永恒的爱”,那么他关注的是电影故事的主题,而另外一个人可能标注的是“震撼”,那么他可能指电影场景或者音乐很震撼;而电子商务系统中的商品没有这些语义特性。其次,社会标签系统中不存在像电子商务系统中那样用户显式地对商品进行主观评级,代之以隐式的评级推荐系统推荐,这种隐式的评级机制是利用标注本身的语义特性,或者是用户使用标签的行为模式。尽管社会标签推荐系统和电子商务系统中商品推荐有以上两点不同,但是有一个是相同的:不管是系统的使用者还是服务的提供者,他们最关心的还是所推荐的对象的质量。电子商务系统中所推荐商品质量的好坏会对用户继续使用电子商务系统的愿望产生较大影响,对商品本身的各个方面并不会深刻影响。但是社会标签系统所推荐标签质量的好坏除了会影响用户使用系统的黏性之外,还会影响标签空间中标签的分布,乃至影响整个社会标签系统语义的涌现,进一步影响用户使用标签进行网络信息资源组织与分享的目的。因此,不管是企业界还是学术界,关于提高社会标签推荐系统推荐标签质量的研究是如火如荼。

本书通过广泛阅读相关文献,总结发现有关社会标签推荐系统的研究内容和研究方向主要包括:

1. 标签推荐技术研究:即推荐算法的研究。当前,社会标签推荐系统所采用的推荐技术主要是借鉴电子商务网站系统的推荐技术,主要包括基于内容的推荐技术^[4] ^[5—7]、基于协同过滤的推荐技术^[8—11]、基于关联规则的推荐^[12,13]和结合前两种技术的混合技术^[14]。然而,如前所述,由于电子商务系统和社会标签推荐系统在推荐对象和评价机制上的不同,使得直接移植电子商务系统中商品推荐的技术到社会标签推荐系统可能会丢失标签所富含的语义和其他客观特征。所以,必须改进电子商务系统中的推荐技术或者结合社会标签系统的新特性发现新的算法来进行社会标签的推荐。

2. 推荐标签质量研究:即推荐对象的研究。因此,标签的好与坏直接影响了标签的各种应用。比如,标签如果存在大量的错拼误写或者歧义缺陷,就会影响标签在检索时结果的召回率和准确率。标签质量方面的研究主要体现在标签与资源的相关性程度、标签推荐的效率、标签推荐结果的评价方法等各个方面。

3. 标签推荐深层机理研究:即人的研究。尽管对社会标签系统中标签推荐的研究越来越多,但是基于用户使用社会标签系统的动机和如何进行标注这个角度来进行推荐的研究却非常罕见。一般认为,用户使用社会标签系统进行标注时,往往不会花费太多精力去深思熟虑应该选择哪些词来作为标签,更多地会用自己或者他人以前使用过的标签进行标注^[15,16]。因此,很多社会标签系统在设计系统时,为了提高用户对系统的可用性和黏性,在系统人机交互界面上提供选择列表以供选择。这个列表可能是用户自己使用过的历史标签,或者是所有其他标注者对该网络信息资源所使用过的历史标签。但对于更深层次的原因没有进行挖掘,比如为什么用户会喜欢选择自己的历史标签,为什么会喜欢使用别人用过的标签,而不是自己创建新的标签。所以对用户标注的深层机理也是目前社会标签系统研究的主要内容之一。

1.3 本书研究的主要内容及成果

本书将研究主题限制在标签的推荐技术上,即针对一个网络信息资源,采用一种合理的算法为用户推荐一些标签来标注这个资源。主要研究内容和成果包括:

1. 针对内容特征的粒度问题, 本书提出了分别使用网络信息资源的内容、已经存在的用户所标注的标签集合以及两者的混合作为建模对象, 从细粒度的词特征和粗粒度的隐含话题两种粒度来建模社会标签。在细粒度的词特征上, 使用经典的统计语言模型来建模, 在隐含话题粒度上, 使用 LDA 模型、TT 模型、UTT 模型分别建模, 最后, 混合两种粒度来建模。通过对 Bibsonomy 的数据集上的考察, 结果表明并不是引入特征越多的模型推荐的效果就越好。相反, 本书中不管是在网络信息资源本身还是已经存在的标签集合上, 统计语言模型具有最好的性能, 引入特征少的隐含话题模型推荐的性能要比引入特征多的隐含话题模型好。

2. 为了解决标签天生的歧义和多义缺陷带来的标签在不同的话题上的重要程度有很大差别, 进而影响标签在检索和推荐应用上的召回率和准确率的问题, 本书在进行标签的重要性排序时, 提出了基于话题敏感的标签排序模型。模型假设每个标签都有很多话题, 并且标签之间依话题的依赖程度存在转移。因此, 该模型首先构建了标签空间中的标签概率转移图, 然后在该图上运行随机游走模型, 这样可以计算出每个标签在话题上的重要性排序。论文将该模型应用于 Delicious 上社会标签的推荐工作, 实验结果表明基于话题敏感的社会标签排序相对于其他方法来说有着更好的效果和更广泛的实用性。

3. 为了给用户推荐更符合用户意图的社会标签, 提高用户对社会标签系统的使用黏性, 提高社会标签在信息资源组织、检索、分类等方面的应用, 更有利于社会标签空间的收敛。本书深入分析标签空间, 提出用 5 种度量方法来度量用户标注的动机倾向性, 从而将用户根据动机倾向性的不同分为描述动机用户和分类动机用户。在讨论了 5 种度量的有效性之后, 根据用户标注历史, 判定用户的动机倾向, 同时计算信息资源的动机倾向。在此基础上, 提出了基于用户动机的社会标签推荐模型。并在两种不同性质的数据集上进行了实验, 实验结果表明基于用户动机倾向性的社会标签推荐模型比其他基准方法具有更好的推荐性能。

1.4 本书的组织结构

全书共分为七章, 各章节按照如下方式进行组织:

第一章主要介绍了本书的研究背景和社会标签推荐系统的主要研究内容, 概述了本论文主要的研究工作及取得的主要研究成果。最后, 给出了本书的整

体组织结构。

第二章主要对本书研究技术的应用背景所涉及的相关内容进行了介绍。阐述了标签、社会标签和社会标签系统的相关问题,主要包括社会标签推荐系统推荐的对象、社会标签标注方法、可视化问题及标签利用等问题。

第三章主要对本书研究的推荐技术及相关理论进行了梳理。对传统推荐技术包括基于内容的推荐技术、基于协同过滤的推荐技术和混合推荐技术的基本思想、推荐过程、推荐算法及优缺点进行了介绍;同时对推荐效果的评价方法和评价指标进行了讨论;最后给出了社会标签推荐系统所面临的问题。

第四章分析了网络信息资源内容上的不同粒度,提出基于细粒度的词特征和粗粒度的隐含话题的社会标签推荐模型,并将此模型应用在实际的社会标签系统 Bibsonomy 中,分析了不同粒度社会标签推荐模型的效果。

第五章在分析了社会标签所具有的话题敏感性之后,提出了基于话题敏感的社会标签推荐模型,并将此模型应用在社会标签系统 Delicious 的标签推荐中。并对推荐的结果进行了评价分析。

第六章在建模用户动机倾向性的基础之上,提出了基于用户动机倾向性的推荐模型,并在不同性质的数据集上进行了实验分析。

第七章对全书工作进行了回顾和总结,指出本书主要的创新点和存在的不足,并对下一步的研究工作进行了简要的探讨和展望。

2 标签及社会标签系统

2.1 标签和社会标签

标注(Annotation)是阅读者对文档内容的一种理解、观点和概括,或者是对文档内容所发表的观点、态度并提出建议的一种途径。早在宋代就有了标注的利用,刘敞、刘攽、刘奉世对《汉书》的文句进行了批注解释,形成了考据史学名著《三刘汉书标注》。不过,也有学者把标注理解为“做记号”,记号可以是与它所标示的内容完全独立的符号系统,也可以是与所标示的内容存在直接关系的词语,但至少可据此理解“标注”具有标示作用,可提高查找的便利性,具有一定的标引功能。

在 Web 2.0 中,用户不仅可以通过豆瓣来分享图书、通过土豆来共赏视频,而且可以通过博客发表博文、通过 Flickr 发布照片、通过 Youtube 上传视频等方式来创造内容。用户的这些活动有一个共同的特征,即用户会自由地选择一些词(Term)或短语(Phrase)来表达自己对这些网络信息资源的某种观点、态度或理解。本书中,我们称用户的这种行为为“标注”(tagging),用户所选择的词或短语为“标签”(tag),提供标注行为的系统为“社会标签系统”(Social Tagging System)。从某个用户来看,标注过程本身非常简单,是个体行为,但这种个体行为的结果—标签具有重要的意义,正是这些标签连接起了用户和资源。一旦标签被 Internet 汇集在一起,形成社会标签(Social Tags),这个时候,汇集在一起的社会标签可以看作社会大众对某个目标资源的理解、观点或者看法。如图 2-1 所示,截至 2014 年 7 月 16 日,豆瓣用户使用“全球化”标注《世界是平的》这本书 4093 人次,使用“经济”标注 3633 人次,由此可以看出,大众认为《世界是平的》这本书是有关经济全球化的一本好书。



图 2—1 “豆瓣读书”上《世界是平的》的图书页面

以 Delicious、Flickr 网站的普遍使用为标志, 提供标签服务的社会标签系统成为众多网络用户在 Web2.0 上的主要应用之一。在这些网站上, 用户从自己的词表中选用简短的字词对自己感兴趣的网页进行标注, 标注结果保存在标签系统中, 不受时间和地理位置的约束, 随时检索和管理。许多专家认为, 这种标签方式的出现使得标签系统能更好地为用户组织资源、浏览资源提供帮助, 极大地方便了用户对个人网络资源的管理和组织。

通过用户对网络信息资源自由标注标签的方式, 在用户和资源之间就有了社会标签这样一个桥梁, 进而形成了用户、资源与标签间的三元组(用户—标签—资源)。由于这个三元组的关系是公开的, 每个用户都可以看到自己所加注的标签与资源, 并通过在网络上游走, 可以看到整个网络上用户、标签与资源之间的关系。

近年来, 在 SIGIR、WWW、KDD 以及其他会议和期刊上有越来越多的研究工作开始关注社会标签。截至目前的研究中, 主要对社会标签的理念和优缺点[17, 18]、统计特性[15][19]、社会标签改善搜索性能[20—22]等进行了研究。文献[17]详细阐释了社会标签系统的理念, 认为社会标签是一种元数据, 文献[18]认为社会标签由<User, URL, Tags>三元组组成(如图 2—2)。Ciro Cattuto 等抓取 Delicious 和 BibSonomy 两个网站的大量数据集, 分析了由三元组构成的超