基于社会化标注的个性化推荐算法研究

魏建良/著

科学出版社

基于社会化标注的个性化推荐算法研究

魏建良 著

科学出版社

北京

内 容 简 介

随着信息社会与数字经济时代的全面到来，越来越多的用户成为互联网信息内容的创造者，网络信息过载也日益严重。在此条件下，如何有效地过滤与选择信息成为时代性的挑战。标签作为一种用户视角的资源特征表述方式，成为个性化信息推荐研究重要的数据来源。本书首先对标签相关文献进行了系统回顾，然后以标签及社会化标注为切入点，应用派系聚类法和向量模型法，从用户间协同、用户多兴趣两个角度构建了若干个性化推荐算法。并在此基础上，结合 WordNet 进一步提出了面向语义优化的改进推荐算法。实验表明，本书所提出的算法具有更好的推荐效果。

本书适合对个性化推荐有兴趣的研究者阅读，也可作为互联网公司技术部门工作者的参考用书。

图书在版编目（CIP）数据

基于社会化标注的个性化推荐算法研究 / 魏建良著. —北京：科学出版社，2019.3

ISBN 978-7-03-060857-4

Ⅰ. ①基… Ⅱ. ①魏… Ⅲ. ①聚类分析－分析方法－研究 Ⅳ. ①O212.4-34

中国版本图书馆 CIP 数据核字（2019）第 049637 号

责任编辑：陶 璇 / 责任校对：杨 赛
责任印制：张 伟 / 封面设计：无极书装

科学出版社出版
北京东黄城根北街 16 号
邮政编码：100717
http：//www.sciencep.com

北京虎彩文化传播有限公司印刷
科学出版社发行 各地新华书店经销
*
2019 年 3 月第 一 版 开本：720 × 1000 1/16
2019 年 3 月第一次印刷 印张：11 1/4
字数：225 000

定价：90.00 元

（如有印装质量问题，我社负责调换）

前　言

当前，随着 Web2.0 中新应用的日渐普及，传统互联网的信息发布模式受到了越来越多的挑战。用户逐渐成为信息内容的生产者和组织者，用户生成和组织内容也成为互联网信息研究中的一个热点。其中，社会化标注（social tagging）是众多研究关注的焦点。允许不同用户对任何资源添加不受词表束缚的，同时又是基于自身理解的标签，促成了标签的流行与社会化标注的发展。可以说，社会化标注构建起用户与资源间全新的关联网络，并为信息资源的推荐提供了新的思路。具体表现如下：一方面，标签是用户对信息资源的理解，是用户偏好的主动表达和真实体现；另一方面，大量用户标注行为中所浮现出的主流标签，是主流用户对资源的理解，更为贴切和恰当地代表了信息资源的内容特征。

本书首先在对社会化标注的基本理论进行简单回顾的基础上，详细地考察目前基于社会化标注进行信息推荐方面的相关研究进展，主要是对其中的排序和聚类算法、用户模型构建及在标签语义改进等方面的成果进行梳理。然后，结合现有研究在用户模型构建方面的不足，提出两种新模型：用户协同模型和多兴趣模型。其中，用户协同模型是为了弥补用户在标注过程中所存在的偏差行为，通过吸收资源中的主流标签到用户模型，来达到偏差矫正的目的，使得存在标注偏差的用户与主流用户的认识达成一致，从而保证信息推荐的质量。同时，由于用户往往是多兴趣的，将多个兴趣主题的标签混合放置于同一向量模型中，极有可能产生标签间语境混乱，进而影响推荐的质量。为此，本书提出子兴趣的概念，通过派系过滤法（cluster percolation method，CPM）聚类算法对用户标注的标签和资源分别加以处理，识别出用户的多个子兴趣并确定其兴趣度。最终的用户模型表示为多个子兴趣集合的形式。

在此基础上，本书对用户协同模型和多兴趣模型算法加以模拟实现，构建相关的模型并给出针对具体模型的推荐。之后，结合用户参与评分的方法对算法的推荐效果进行评价。研究发现，基于用户协同模型的算法略优于现有的基于用户自身标签的算法，而基于多兴趣模型的算法则明显好于上述两种算法，原因可能是多兴趣模型中的子兴趣保持了资源主题的单一性，从而有助于找到更为相关的资源。

最后，鉴于推荐算法中存在的标签同义和多义问题，本书提出相关的解决策略，并进行实证分析。对于同义标签的处理，主要思路是借助 WordNet 找出与目

标标签有相同含义的词，并将其吸收到推荐模型中。而对多义标签，则是结合 CPM 对标签进行聚类分析以识别多义标签。在此基础上，给出多义标签的邻居标签集并计算其与所在资源模型间的相似性，进而确定多义标签的具体含义，并将该含义对应的邻居标签集吸纳到推荐算法中。通过增加这种额外的信息，来克服同义和多义问题对推荐的影响。

目　录

第一篇　基　础　篇

第二篇　基础算法篇

第三篇　语义优化篇

第四篇　结　论　篇

第一篇 基 础 篇

本篇首先给出本书研究背景与意义，社会化标注可以为个性化信息推荐带来新的思路，并有助于提高信息推荐的质量。其次，对社会化标注的产生与内涵、机制、优势与不足等进行简要的介绍，重点对社会化标注的研究现状进行系统的梳理，包括标签的性质、社会化标注系统的相关模型、标签对推荐模型的意义、基于社会化标注的聚类和排序算法、标签的推荐、基于社会化标注的个性化算法、标签的语义分析等方面。

在此基础上，发现用户模型构建不尽合理和推荐算法中对标签同义多义问题考虑的缺失是现有研究中存在的主要不足。基于此本书提出主要关注内容，包括用户协同模型和多兴趣模型的构建，以及推荐算法中对标签同义多义问题的改进。

第1章 绪 论

在 Web2.0 的环境下，社会化标注服务出现伊始，就在产业界得到了广泛应用，出现了书签（如 Delicious）、照片（如 Flickr）、视频（如 YouTube）、书籍（如 LibraryThing）、音乐（如 Last.fm）、引用（如 Connotea、CiteULike）、博客（如 Technorati）等众多新的应用与体验。社会化标注允许任意用户对感兴趣的网络资源进行基于自身理解的无约束标注，并且所有用户的标注都互为可见。这种开放、共享的模式及反映用户真实观点与偏好的标注为网络信息资源的组织和共享带来了一种全新的理念，它是一种大众智慧的体现，具有潜在的发展优势。

1.1 研究背景及意义

在信息技术的推动下，互联网在最近十几年经历了异常迅猛的发展。无论是网络的用户、带宽，还是内容的丰富程度，都已经得到了极大的提升。可以说，网络已经成为当前人们生活中不可缺少的一部分。特别是互联网上丰富的信息资源，为人们的学习、工作和生活都提供了很大的便利。据相关机构统计，截至 2016 年 3 月，全球至少有 46.6 亿个在线网页。但与此同时，人们在如何充分和有效利用互联网信息资源的问题上碰到了困难。尤其是在海量的信息面前，人们往往会变得不知所措、无从下手。

鉴于此，在互联网实践者和研究者的努力下，出现了分类目录、搜寻引擎，以及推荐系统等信息过滤技术工具，这些技术的应用，大大提高了用户查找和获取合适信息的能力。尤其是个性化推荐技术的应用，通过主动向用户推送其所感兴趣的信息，从根本上改变了用户获取信息的模式，提高了用户获取信息的效率。但是，传统的个性化推荐技术存在一些固有的不足，其中最根本的问题是不能及时、准确地追踪或描述用户的兴趣模型，从而影响信息推荐的质量。而伴随着 Web2.0 的兴起，互联网中信息资源的生产和组织方式都发生了一些新的变化，尤其是凸现了用户参与的重要性。这些变化，一方面，使得信息的组织更加面临挑战；另一方面，为推荐系统的发展提供了新的思路。

2003 年以来，互联网发生了一系列新的变革：从网络博客（weblog）出现，到博客（blog）、社会化标签、内容聚合、对等网（peer-to-peer，P2P）、维基（Wiki）、

阿贾克斯（Ajax）、Web 服务、社会化网络软件等技术和理念的相继产生，形成了互联网应用的新一代发展趋势。特别是在其中的社会化标签领域，涌现出了许多标注服务与体验网站，包括书签、照片、视频、书籍、音乐、引用、博客等多个主题[1]。2005 年 9 月，Tim O'Reilly 较为全面地对 Web2.0 这一概念进行了阐述，作为对互联网环境中出现的上述新的技术理念与商业模式的概括。

2007～2009 年，标注书签网站 Delicious 和图片共享网站 Flickr 成为众多新兴的 Web2.0 服务中的先行者与佼佼者，标签（tag）也成为网络的一种新时尚。据美国皮尤研究中心（Pew Research Center）在 2007 年 1 月发布的调查报告，有 28% 的美国互联网用户对在线的照片、新闻或者博客进行了标注和归类，一天中平均有 7%的用户称其对在线内容进行了标注或归类[2]。而根据著名的博客搜索引擎 Technorati 的统计，2006 年 3 月有 47%的博客日志运用了标签进行标注，并以平均每天 56 万篇的速度增加[3]。最近几年，标签更是在大范围内得到了应用，包括 Flickr、MovieLens、豆瓣等社交平台均将用户添加的标签作为信息生成的重要方式，仅一个平台中积累的用户标签的数量就突破几十亿个，许多平台甚至将自动标签系统作为了信息分类的重要支撑。

这种互联网的新变革给用户带来了截然不同于传统模式的体验。主要表现为：信息发布由传统的集中式发布向分布式发布演变，blog 等逐渐成为互联网信息发布的重要方式。同时，普通用户也开始从纯粹的内容浏览者向内容生产者与组织者身份转变，从互联网的被动接受者和旁观者慢慢转变为内容的主动创造者和管理者[4]。用户不再是被动地接受，而是变得主动且要求自己的需求能够得到个性化的满足[5]。这一点，在美国《时代周刊》（*Time*）2006 年所评选的年度人物上早就已得到了充分体现。2006 年 12 月 17 日，《时代周刊》选择了“YOU”作为年度人物。“YOU”指的是，互联网中的用户在网络社区及协作化行为上都达到了一个前所未有的规模。这是信息社会转变的标志，在网络中专业和非专业人员越来越共同致力于同一问题。

与此同时，传统的推荐技术一直存在部分固有的不足。传统的个性化推荐技术主要包括基于内容过滤的推荐和基于协同过滤的推荐两种方式。在建立用户兴趣模型方面，基于内容过滤的推荐最为典型的做法是抽取信息资源的特征值，如从文本信息中抽取关键词并计算词的权重，进而建立用户模型。基于协同过滤的推荐则是假设对部分相同资源感兴趣的用户其偏好是相似的，或者假设大部分用户对相似项的评分比较相似，则目标用户的评分也与其相似，进而基于这种相似性进行信息推荐。尽管这两种推荐方法为主动推送信息资源提供了重要支撑，但也都存在着一定的缺陷。特别是基于内容过滤的推荐只能抓取信息资源的表面特征，而不能确保对资源中深层或隐含信息的理解，从而无法保证用户兴趣模型的准确性。同时，在基于协同过滤的推荐中如果没有用户积极参与评分，或者评分

信息难以收集，则无法进行有效的推荐。基于协同的过滤还往往面临一定的数据稀疏性和冷启动问题。

在新的网络形势下，传统推荐技术对用户模型的描述与现实之间的差异变得更为尖锐。一方面，用户已成为信息内容的重要生产者，信息的生产与发布不再是“权威者”的专利。这推动了信息资源在类型、内容及表达方式方面的极大丰富，但对于无偏差地从信息资源中提取客观内容变得更为困难了，进而导致通过内容过滤方法得到的用户兴趣模型有可能更大程度地偏离客观现实。另一方面，大量普通用户成为信息的生产者，推动了互联网中信息量呈指数式膨胀，并且这些信息是分散分布的。如果应用协同过滤技术对这些信息进行推荐，不仅面临着难以收集分散的评分数据的难题，还会受到更为严重的数据稀疏性与冷启动问题的困扰。

然而，Web2.0 带来的并不都是混乱，新的服务和应用的出现与流行必然有其合理性。社会化标注就为这种分布式的内容生成形式提供了有效的管理手段，并为推荐技术提供了新的思路。在介绍社会化标签之前，先讨论一下标签。标签类似于关键词，它是用户对发布的信息资源所添加的一种基于关键词的描述。但它与关键词不同的是，标签的添加不存在权限和词汇形式的限制。不仅信息发布者可以对信息资源添加标签，信息浏览者也可以对信息资源添加标签。任何用户可以对任意信息资源添加基于自身理解的不同标签。当有大量用户对大量的信息资源添加了标签，并相互之间形成共享和交互时，标签就具有了社会性，成为社会化标签。进一步地，将用户添加标签的行为称为标注，大量用户对大量资源进行的共享性标注行为，也就构成了社会化标注。概括地说，社会化标签具有两个明显的特征：一是在对资源添加标签的过程中，用户不需要遵循任何事先既定的分类法或者词表，所添加的标签可以完全是基于自身的理解的；二是每个用户的标注行为是开放和共享的，社会化标注将用户、标签和资源三者内在地联系了起来，形成了重要的关联网络。

与此同时，用户在社会化标注中往往倾向于对相似的资源添加类似的标签，因此，通过这些标签就可以找到相关联的资源，这在一定意义上形成了信息资源的分类法。信息构建专家 Thomas Vander Wal 将这种基于互联网的社会环境，并由大众用户产生的信息分类组织方式命名为 folksonomy，译为大众分类法。有研究者认为[6, 7]，由于大众分类法的产生，用户可以使用自己的词汇对信息资源进行标注，方便了资源再次查找和使用。更为重要的是，相同的标签能够聚合整个资源空间中的所有相似内容，实现资源的共享。基于标签的浏览更能让用户获得意外的发现，用户在浏览的过程中能够找到与自己拥有相同兴趣的用户，进而发现这些用户所标注的其他潜在兴趣资源。可以说，大众分类法的形成和发展具有明显的社会化的性质。

一般而言，标签是用户对资源内容的高度概括，蕴含了用户对资源特征的分析与重新表达。由于用户所具备的知识和经验，往往能够比基于词频分析的机器算法更能把握信息内容的中心词和重点，所以即使标签所用的词语在资源中的频率较低，也可能比那些词频较高的词汇更能反映信息资源的本质特征。对于标签的作用，有研究者认为，标签是一种由用户产生的元数据，但它又与以往由专家或文章作者产生的元数据不同，它能够直接、迅速反映用户的词汇和需求及其变化[6]。目前，Delicious 等标签服务网站已有大量的标签，足够让人们去发现隐藏在其中的模式。对单个资源来说，其标签的分布较为稳定，频率最高的那部分标签比例较小且稳定。根据一项研究的数据，10%的最流行标签覆盖了所有网页资源的84.3%[8]。尽管不能避免不同用户在标注过程中会存在不同认知的问题，但社会化标注的一致性会随着信息资源在互联网上流行性和收藏数的上升而提高[9]。此外，社会化标注中不用进行评分数据的收集，用户在进行标注的同时即明确了目标资源的标签信息。同时，信息内容在发布时就进行了标注，面临冷启动问题的可能性很小。

可以说，社会化标注中的标签不仅是对资源特征的良好描述，还能较好地代表用户的兴趣偏好，用户往往是对自身所感兴趣的资源进行标注。因此，通过社会化标注，建立相应的个性化推荐算法，是突破困扰传统推荐算法所面临问题的一个新思路。其潜在优势包括：第一，在社会化标注环境下，用户兴趣信息从原有的被动收集变为主动表达，其标注的资源都与其兴趣高度相关，因此，有可能建立更为准确、详细的，能反映用户真正偏好的用户兴趣模型[10]。第二，社会化标注系统中，无论是流行资源还是新添加的资源，都会有一定数量的标签存在。这些标签不仅是对资源的描述，还可以将目标资源和用户与其他资源和用户联系起来，从而有可能从根本上解决信息推荐中的冷启动问题。第三，基于社会化标注的推荐算法，通过大量“用户—标签—资源”间的关系网络，有机地整合了基于内容过滤与协同过滤算法的思想，具有内在的固有优势。

1.2　社会化标注的相关理论

本节对社会化标注的基本理论进行介绍，主要包括社会化标注产生与内涵、内在机制、优势与不足等。

1.2.1　社会化标注产生与内涵

在讨论社会化标注之前，首先了解一下标签的来源。标签并不是最近出现的

新鲜事物，特别对于图书馆馆员、编目者和专业分类人员而言，标签的使用已有较长历史，但是其所用的标签是受控的，而且没有体现出社会性。而本书所指的社会化标注，是指在 Tim O'Reilly 于 2004 年首次提出 Web2.0 概念后，大量用户通过添加关键词到信息资源并体现 Web2.0 核心思想的行为，同时可以实现对资源的自动分类，这是一种无须可控词汇但有效的主观索引。在这种新的方式下，每一个用户都在进行标注，而不再是一小部分专家进行标注，标签走向了公开化，并形成共享[4]。

最早对社会化标注的关注起源于 1997 年 Keller 等建议通过协同方法加强网络浏览器的书签功能[11]。之后，Bry 和 Wagner 也开展了一项类似的研究[12]。受此启发，Joshua Schachter 在 2003 年年底开始了第一个社会化标注服务网站，也就是现在的 Delicious。随后，许多不同主题的社会化标注网站与系统开始建立。

社会化标注在英文里有着较多类似的表述，如 social bookmarking、social tagging、social annotation、collaboration tagging、folksonomy、social classification、social indexing 等。在这些概念中，除 social bookmarking 的对象为书签外，其余的概念都表达了类似的内涵，其对象不仅可以包括书签，还涵盖了图片、视频、参考文献、博客、图书等众多互联网资源。从理论而言，所有的网络资源都可以用社会化标签进行标注。

在上述表述中，social bookmarking、social tagging、social annotation 和 collaboration tagging 侧重描述的是社会化标注的行为，强调标注行为中体现的协同性。而 folksonomy、social classification、social indexing 则侧重于对标注结果的描述，强调的是标注对资源分类所产生的影响。其中，最有代表性的是 folksonomy。folksonomy 是由 Thomas Vander Wal 在讨论 Flickr 和 Delicious 所发展的信息架构时，将 folks 和 taxonomy 组合而成的新词[13]。他称这样的架构是由下而上的社会性分类（bottom-up social classification）法。其中，folks 本意指一般人、大众等，而-sonomy 则是由 taxonomy 演变而来，表示一种有系统的、专门的学科知识。将两者合而为一的意义是：由大众所产生的一种分类知识。另外一种看待 folksonomy 的方式是，如果将 taxonomy 看作分类系统的全部，那么本体（ontology）是为分类系统的架构指定正式的名称，并使这些名字能合适地描述架构。而 folksonomy 则有点像人们按照自己的方式对这些架构命名，然后采用最为流行的那个名字。因此，folksonomy 是一种综合的行为，而不只是创造标签[14]。

根据维基百科（Wikipedia）的定义，大众分类是指协同创造和管理标签，实现对内容进行标注和归类的方法与实践。其中突出说明三点：第一，标签是由个人用户所创造的，而且标签的选词是根据用户对信息资源的理解，在形式和内容上不受已有词表的限制，即大众分类是由用户所建立的无结构的扁平词表[15]；第

二，标签和标注的环境是基于共享和开放的，任何用户可以对任何资源进行标注，且标签可以相互可见；第三，大量个人用户的标注行为通过碰撞与融合，形成了社会性，资源实现了标签条件下的自动归类。可以认为，大众分类是一种基于互联网的信息检索方法，是由用户合作创建，并由开放端标签所组成，将网页、在线图片、网页链接等众多的网络内容加以分类的方法[15]。

可以看出，在社会化标注中，主要的对象有三个，即用户、资源和标签[16]。用户包括资源的创建者、标注者或使用者。在社会化标注系统中，绝大部分的用户是互联网的普通使用者。资源是指存在于互联网中的各种类型的信息，如网页、文献、博客、图片、视音频等资源。标签是指用户所选择的进行标注的词汇。尽管这些词汇源自用户的语言，通常具有随意性和多样性的特点。但是它们是基于用户对事物的理解，往往更为真实地体现了用户的需求，并能迅速反映用户需求的变化。在社会化标注中，标签不仅能够聚合类似的资源，推动用户社群的形成，更为重要的是建立了一条连接用户与资源的纽带。正如 Bateman 所指出的，社会化标注不是添加关键词的简单行为，它是大量用户对事物特定看法的词汇。因此，通过观察社会化标注行为，包括用户、标签和资源及它们之间的联系，就可以得到比简单的关键词更为丰富的视角[17]。

在图 1.1[18]所示的标注系统模型中，将社会化标注看作一个系统，其中的元素就是用户、资源和标签，通过标注行为建立了三者间的联系。具体而言，用户通过在资源中添加标签而建立了用户与资源间的关系（图中实线），这种关系是社会化标注系统运行的基础。相同的标签被运用于不同的资源，表明在用户的认知中这些资源可能具有某种共性。而使用相同标签的用户，特别是当他们对特定的资源使用相同标签时，表明他们的知识体系有某种重合，因而标签能够帮助用户

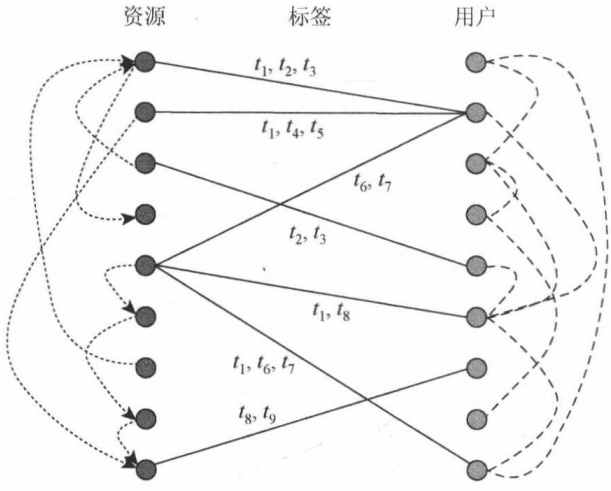

图 1.1　标注系统模型

发现具有“共同兴趣”的群体。同样，即使不同用户对同一资源使用不同的标签，这些标签之间也可能存在同义或相关的关系。此外，通过标签或者用户，资源之间也可以建立相互间的联系（图中左侧虚线）。同时，通过标签和资源，不同用户间也可能存在关联。当然，这种联系也可能是独立的，即用户在现实中就属于同一集团，或者加入了同一个社区（图中右侧虚线）。

应当意识到的是，并不是所有的标注行为和个性词表都能使得一个共同的大众分类出现，因此有必要找出广义和狭义大众分类的区别[15]。事实上，Wal 给出了两种类型的大众分类，即广义大众分类（broad folksonomy）和狭义大众分类（narrow folksonomy）两种[13]。广义大众分类最为典型的代表是 Delicious，如图 1.2 所示。在广义大众分类中，每个用户可以对任一资源添加标签，这样的系统中常常是大量用户对同一个资源进行标注[19]。这些用户一般都具有不同的知识结构和兴趣领域，他们所标注的标签都是基于个人的特定背景。一般而言，使用类似的标签对相同资源进行标注往往代表了这些用户具有某种相同的偏好，这种类似的标签体现了用户对资源的多个角度的描述，从而加强了资源的可检索性。同时，大量用户根据其所标注的标签就可以分化为一个个小团体，从而为寻找相似用户提供了途径[15, 20]。广义大众分类的缺点是采用大量不同的关键词描述事物，以及由此而产生的分散性，会使寻找特定信息变得困难[20]。

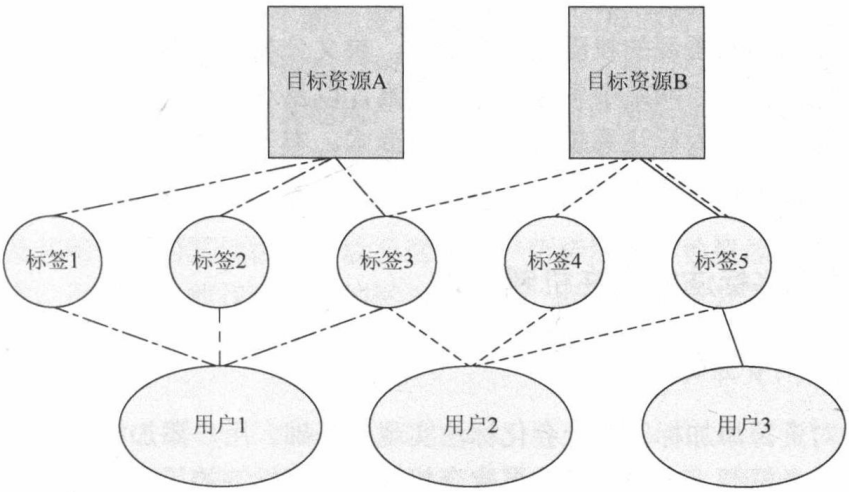

图 1.2 广义大众分类

狭义大众分类的典型代表是 Flickr，如图 1.3 所示。在 Flickr 中，虽然图片和标签可以被所有用户查看，但对某一图片只有其所有者和该所有者定义的“好友”才拥有添加标签的权力[19]。因此，在这样的系统中，对单个资源一般都只标注了较少的标签，但每个标签所对应的资源数相对较多，而且标签所用的词汇往往很

相似。因此，尽管缺少了广义分类法中词汇的丰富性和多样性，但在狭义大众分类法中利用单个关键词就能较为准确地找到相关资源[20]。狭义分类的缺点是对资源进行描述的标签较少，且缺少交互性。一般只能在社区的小群体中运行较好，在大的系统中的效率并不高[15]。

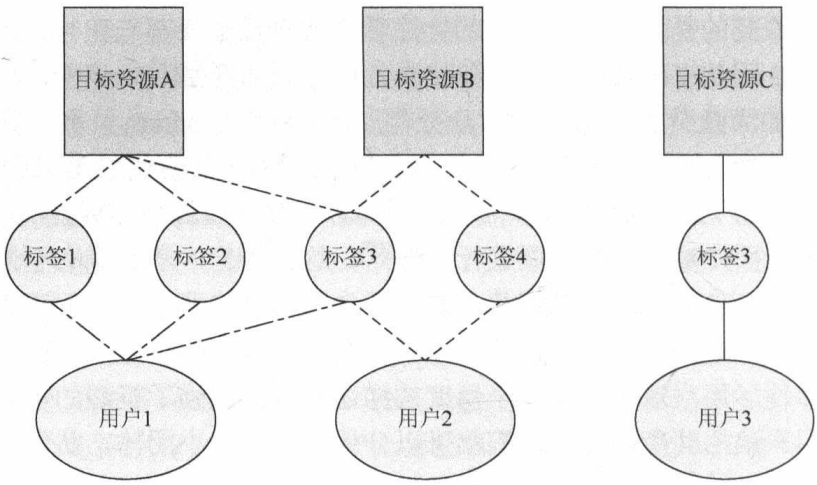

图 1.3　狭义大众分类

应当说，标注的权限和用户规模大小是广义和狭义大众分类法的根本区别。根据用户规模和对资源的权限设置的变化，狭义分类法可以逐渐向广义转化。同时，随着社会化标注网站中资源的日益丰富，以及对资源共享的追求，广义的大众分类将越来越成为标注系统发展的主流形式。本书所讨论的社会化标注指的是广义大众分类中的标注行为。

1.2.2　社会化标注的内在机制

1. 用户对资源的标注

用户对资源添加标签是社会化标注实现的基础。用户添加标签的最初动机是管理个人信息资源、方便资源的再次查找和使用。标签的用词代表了用户对资源的描述或理解。标签可以是用户认为对自身有意义的任何词，或者是对资源在抽象意义上的理解或概括，无论该词在资源中是否出现。

用户对资源所添加的标签对于信息推荐而言意义显著：一方面，标签真实地体现了用户对资源的理解和概括，具有较高的可信性；另一方面，标签也充分折射出了用户兴趣，标签的集合往往可以作为用户兴趣模型的一种表达。

此外，很多的实证研究指出，无论是用户对标签的使用，还是在单个资源中