情报学报 ISSN 1000-0135
第35卷第5期549-560，2016年5月

JOURNAL OF THE CHINA SOCIETY FOR SCIENTIFIC AND TECHNICAL INFORMATION ISSN 1000-0135
Vol. 35 No. 5, 549-560 May 2016

doi:10.3772/j.issn.1000-0135.2016.005.011

社会化标注系统中的个性化信息推荐研究[1)]

熊回香 杨雪萍

（华中师范大学 信息管理学院，武汉 430079）

摘要 在多媒体网络平台中，不仅社交网站允许用户自由发布资源和添加标签，越来越多的资源共享系统也开放给用户对资源、标签的组织管理权限。本文在分析了社会化标注系统的利弊后，采用推荐技术解决社会化标注系统中资源获取困难的问题，构建了基于社会化标注系统的个性化信息推荐模型，提出了从资源-标签-用户三个维度分别建立推荐组件，进而重组推荐资源集合实现对用户的个性化兴趣预测算法，并选取豆瓣网上的实例数据验证了算法的可行性和有效性。

关键词 社会化标注 标签 资源推荐 聚类分析 协同过滤

Personalized Information Recommendation Research Based On Combined Condition In Folksonomies

Xiong Huixiang and Yang Xueping

(*Department of Information Management, Central China Normal University, Wuhan* 430079)

Abstract On the Multi-media platform of network, apart from the social websites allow users to release and add tags, more and more information resource sharing system open to all people permissions of organizing and managing items and tags. This paper analyzed the advantages and disadvantages, applying recommendation technology to solve the problem which is hard to obtain useful information. This paper constructed a personalized information recommendation model which is based on the folksonomy system. It proposed to design three recommendation components based on three dimentions of resource, tag and user respectively, then combined the three candidate sets in order to realize the user interests prediction. It took Douban Reading as an example to describe the proposed model and displayed the results of personalized information.

Keywords folksonomy, tag, information recommendation, cluster analysis, CF

1 前 言

社会化标注又叫协同标注、大众分类等，是指用户可以自己创建标签来对网络环境中的虚拟内容（包括博客、图片、视频、音频、用户资料等）添加注释[1~3]。伴随互联网技术的高速更新发展，涌现出了大批允许用户自由创建内容的社会化标注系统，国外有 Delicious.us（共享标签）、Flickr（共享图片）、Last FM（共享音乐）、YouTube（共享视频）等[4]，国内有豆瓣、微博、花瓣网等。在互联网早期，网页设计人员使用网页元标签关键词，告诉搜索引擎有关网页的内容，而社会化标注系统为用户在面对信息过载的网络环境中，提供了寻找有用资源的新方法。但是，大众在自由创建标签时的完全无监督

特性产生的问题（如标签模糊、标签冗余、标签歧义等），一方面会降低内容标引和检索的有效性[5]；另一方面，更多的有效资源因很少受到访问

收稿日期：2015年12月14日

作者简介：**熊回香**，女，1966年生，教授，博士，主要研究方向：网络信息组织与检索，E-mail：hxxiong@mail.ccnu.edu.cn。**杨雪萍**，女，1992年生，硕士研究生，主要研究方向：网络信息组织与检索。

1） 国家社会科学基金项目“大众分类中标签间语义关系挖掘研究”（批准号：12BTQ038）。

而被沉没,反而阻碍了用户操作,事实上,帮助用户获得这些"暗信息"[6]比热门资源更能提高用户体验。因此,应用个性化推荐技术[7]以弥补社会化标注系统中的检索结果不精确、信息获取困难等问题,成为研究和关注的重点[8]。关于推荐对象,目前国内对社会化标注系统的推荐研究主要集中于标签推荐及相似用户发现两个领域,关于资源推荐的研究还很少[9~11],而获取资源才是用户检索和使用信息网站的主要目的,故本文选用资源而非标签或用户作为推荐目标。关于推荐方法,大致分为三类:基于内容的(content-based)、协同过滤(CF)及混合(hybrid)推荐方法,本文综合考虑推荐系统运用的技术所具有的冷启动、数据稀疏性、推荐准确稳定性等问题,提出多个组件同时计算相似性,采用提取短文本内容的特征词,计算各组件中对象的相似性进而聚类,并隔一定时间进行更新[7],以缓解计算压力,并在不同组件中选用合适的推荐技术来提高预测准确度,例如基于相似资源的组件中,选用content-based方法,基于相似用户的组件中,选用CF方法。

本文构建的个性化推荐模型应用于已有资源、标签的用户,当用户在检索框中输入检索词后点击结果列表中的目标资源,系统根据用户所选中的资源、用户相关资源(本文指用户读过、在读、想读的图书资源)、以及用户用来标注资源的标签信息计算用户偏好,为用户个性化推荐相关资源。

2 数据来源与推荐框架描述

本文选用国内社会标注网站中用户活跃度较高的豆瓣网站,以该网站的豆瓣读书频道作为实证研究对象,豆瓣网是一个集评论、资源推荐和共同兴趣交友等多种服务为一体的社会化标注网站,定位于帮助你发现喜欢的东西[12]。豆瓣在个性化推荐中做出过很多尝试及更改,主页曾直接调整为豆瓣广播、豆瓣猜,但由于推荐效果不太好,最后将个性化推荐又调整为页面中的模块功能。以豆瓣读书为例,在豆瓣读书首页,按照模块被扫视到的先后,界面由"新书速递"、"热门标签"、"最受关注图书榜"、"畅销图书榜"、"豆瓣猜你可能感兴趣的图书"、"电子图书"、"书评人"七个模块构成,前面的几个模块可视为常规推荐,而曾经最为被看重的个性化推荐模块,即"豆瓣猜你可能感兴趣的图书"被放置到了页面的右下角。大数据环境下,面对社会化标注系统中的信息过载、冷启动等负面问题,解决方法主要从两个主要的互联网技术着手[8]:信息检索和信息推荐,只有结合这两种技术,才有望真正改进检索策略的僵硬化,提高推荐系统的准确度。本文设计的个性化推荐模型整合了社会化标注系统中的三个元素:用户、资源(本文指图书)、标签(热门标签),分别建立了基于相似资源的推荐组件、基于用户检索词的推荐组件及基于相似用户的推荐组件,最后重组三个组件的推荐结果,形成个性化推荐资源集。

2.1 数据收集

为清晰论述提出的个性化推荐过程,需要用到的数据包括:用户、用户相关图书的简介、图书的热门标签作为社会化标注下的图书标签、用户对相关图书的自定义标签作为用户标签。采用手动收集数据的方法,随机选择"豆瓣读书"网站中的4名用户,数据包括该4名用户的所有相关(包括读过、想读、在读)图书及其内容简介、每本图书的热门标签10个、以及用户用于标注图书的标签。

2.2 数据预处理

由于采取的手动收集,基本不存在拼写错误或过分不标准数据,使用张华平博士推出的NLPIR(大数据搜索与挖掘共现平台)对简介内容、用户标签和热门标签完成分词、特征词提取工作,例如"可爱的三毛"、"个人管理"等处理为"三毛"、"管理","中国文学"处理为"中国"和"文学"。最后得到本文提出模型所需的全部描述数据如表1、表2、表3,用户4名,书籍19本,图书特征词共464个,热门标签共114个,用户标签共63个。其中表3由于内容过多,所以简介特征词没有全部显示。

表1 图书资源集

编号	图书
A~S	疯狂Android讲义;灵魂机器的时代;带一本书去巴黎;心理罪:暗河;浪潮之巅;习惯的力量;习惯的力量;亲爱的三毛;山南水北;围城;明朝那些事儿;中国近代史;我们时代的神经症人格;美丽新世界;且听风吟;深入理解计算机系统;我是猫;失落的玫瑰;我承认我不曾历经沧桑;少有人走的路

表2 图书－标签数据集

编号	简介特征词	热门标签
A	讲义;Java;…;认真	android;编程;IT;开发;讲义;计算机;程序;技术;实用;教材
B	人类;机器;…;精彩	人工智能;科普;未来;计算机;科学;计算机科学;科技;机器;IT;社会
C	历史;社会;…;革命	林达;旅行;法国;文化;历史;随笔;巴黎;游记;散文;旅游
D	人质;骗局;…;绝望	雷米;推理;悬疑;小说;犯罪;心理学;心理;罪;中国;现代
E	科技;投资;…;商业	互联网;IT;商业;计算机;历史;管理;投资;Google;云计算;社会
F	习惯;生活;…;幸福	心理学;习惯;完善;管理;力量;心理;效率;思维;励志;心灵
G	书信;生活;…;幽暗	三毛;散文;台湾;随笔;中国;文学;杂文;生活;当代;现代
H	意义;历史;…;生活	韩少功;随笔;当代;文学;生活;散文;文学;中国;杂文;闲书
I	中国;小说;…;深刻	钱钟书;小说;中国;文学;经典;婚姻;现代;围城;人生;爱情
J	政治;小说;…;年代	历史;明朝;明史;中国;小说;有趣;当年;明月;感触;感悟
K	历史;社会;…;深刻	历史;中国;近代史;徐中约;社会;近代;经典;文化;社会学;纪实
L	冲突;文化;…;精彩	心理学;精神分析;霍妮;心理;神经症;社会;哲学;经典;现代;成长
M	社会;幸福;…;快乐	乌托邦;赫胥黎;小说;英国;政治;外国;文学;科幻;新世界;思想
N	文学;小说;…;复杂	村上春树;小说;日本;文学;青春;外国;林少华;孤独;且听风吟;成长
O	程序;技术;…;实用	计算机;系统;计算机科学;编程;程序员;IT;软件;设计;经典;教材
P	社会;文学;…;日本	日本;文学;小说;外国;名著;夏目漱石;猫;讽刺;荒诞;文化
Q	旅行;心灵;…;幸福	土耳其;心灵;小说;外国;文学;觉醒;人生;哲学;心理;神秘
R	时代;文学;…;精彩	蒋方舟;杂文;随笔;青春;成长;中国;文学;感触;散文;社会
S	时代;生活;…;宁静	心理学;成长;心理;励志;心智;人生;爱;心灵;苦难;人性

表3 用户－图书－标签数据集

用户	图书	用户标签	用户	图书	用户标签
Z. Y. T	A	计算机;android;编程	Andrea	L	精神分析;心理学;卡伦·霍妮
	B	人工智能;计算机;科学;科普		M	乌托邦;赫胥黎;小说;政治
	C	林达;旅行;法国;历史;随笔		N	村上春树;小说;日本
	D	犯罪;心理学;中国;小说		K	中国近代史;徐中约;经典;社会
	E	互联网;IT;商业		E	商业;计算机;互联网;历史
	F	心理学;习惯;力量;完善;管理		Q	经典;计算机;科学;操作系统;体系;结构
泥蓝海馨	G	书信;生活;散文;人生;三毛	徐徐又当这信	P	夏目漱石;日本;小说
	B	人工智能;未来;计算机		Q	土耳其;小说;外国;文学
	H	散文;随笔;中国;文学;韩少功		C	旅行 文化 历史 巴黎 林达
	I	钱钟书;小说;文学		H	韩少功;乡土;文学;中国
	J	历史;明朝;当年;明月		G	台湾;书信集;三毛;散文
	K	近代史;反思;历史;徐中约		R	蒋方舟;杂文;散文;随笔
				S	心理学;成长;励志;人生

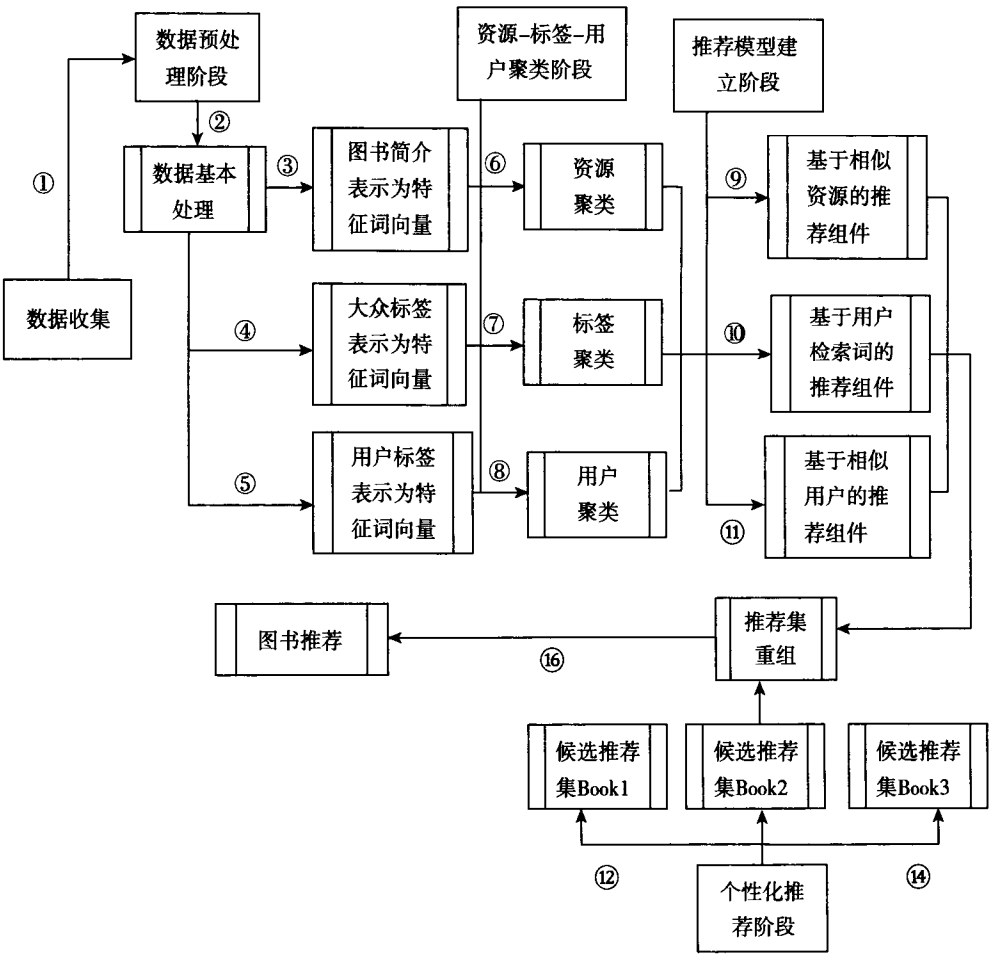

图1 推荐模型的总体框架

2.3 个性化推荐总体框架

当前的推荐系统规模越来越庞大,用户、商品类目数量极多,考虑大部分用户,两两之间的选择重合率很低,造成数据稀疏性[6],而数据让用户与用户(通过标签和资源)、资源与资源(通过资源内容)、标签与标签(通过资源)产生关联显得尤为重要,单纯的考虑用户、资源、标签,数据都会非常稀疏,但如果把信息粗粒度化,综合考虑多种关联信息,数据就会立刻变得稠密。所以,本文建立的个性化推荐系统模型由五个阶段组成:数据收集、数据预处理、资源-标签-用户聚类、推荐模型建立和个性化推荐,融合社会化标注系统中的三元素<用户、资源、标签>分别建立组件推荐资源,然后重组候选推荐资源集,实现最后的个性化推荐目标。总体框架如图1所示,其中,步骤①~⑤在2.2节完成,其中③~⑤表示采用NLPIR提取关键词构成特征词向量。

在本模型论述中,假定用户 U^* 有相关图书资源 N 本,对这 N 本图书共标注了 M 个标签,检索的目标图书为 R^*,以该用户在检索框下输入的检索词为推荐系统的初始点,基于本文提出的个性化推荐模型,为用户推荐感兴趣的图书。本文以名为"泥蓝海馨"的用户,构造检索词"明朝那些事儿",检索目标图书"《明朝那些事儿》"为例。因为本文用到的数据量很少,因此文中提到的 K 值和阈值的相关设定不作讨论,重点论述个性化推荐模型的构建,对过程中涉及的文本处理方法、聚类方法等的选择,是在不影响结果的前提下,以易于实现为标准。

3 数据聚类分析

聚类分析是一种无监督的学习,它将相似的对象归到同一个簇中,它可以被称为全自动分类[13],聚类方法几乎适用于所有对象,簇内的对象越相似,聚类效果越好,对资源、标签、用户实施聚类,是完成后面推荐的前提。

3.1 资源聚类

K-means算法是一种典型的聚类方法,是发现给定数据集的 k 个簇的算法,通过计算数据集中点

与点之间的距离或相似度来实现聚簇，且每个簇的中心采用簇中所含值的均值计算而成[13]。

3.1.1 向量表示

通过将特征词定量化，把用户识别的文本信息转换为计算机可读、可计算的空间向量模型。前面已对图书简介进行了预处理，每本图书都可以用特征词表征，要对资源定量表示，需要给每个特征词赋予权重，考虑到文档长度不同对权重的影响，权重公式需要做归一化处理[10]，将各项权值规范到[0,1]之间，采用改进的 tf-idf 函数进行计算，即公式(1)。

$$w_{ik}=\frac{tf_{ik}\times\log\left(\frac{N}{n_k}+0.01\right)}{\sqrt{\sum_{k\in i}\left[tf_{ik}\times\log\left(\frac{N}{n_k}+0.01\right)\right]^2}};i=1,\cdots,N;k=1,\cdots,m \tag{1}$$

式中：tf_{ik}是第 k 个词在图书 i 简介中的频数，N 是图书总数，n_k是包含第 k 个词的图书数量，w_{ik}是第 i 本图书中词 k 的权重，m 是特征项总数。

根据公式(1)求得所有特征词在每本图书中的权重值，以《疯狂 Android 讲义》为例，如表 4。

3.1.2 相似度计算

采用 VSM 模型中的余弦定理来度量两个图书 d_i 和 d_j 之间的相似性[14]。公式(2)的定义如下：

$$Sim(d_i,d_j)=\frac{d_i\cdot d_j}{|d_i|\times|d_j|}=\frac{\sum_{k=1}^{n}w_{ik}\times w_{jk}}{\sqrt{\left(\sum_{k=1}^{n}w_{ik}^2\right)\left(\sum_{k=1}^{n}w_{jk}^2\right)}}\quad i=1,\cdots,N;j=1,\cdots,N \tag{2}$$

由于对特征词的权重进行了归一化，因此，$|di|=1$。所以文本间的相似度简化为公式(3)：

$$Sim(d_i,d_j)=d_i\cdot d_j=\sum_{k=1}^{n}w_{ik}\times w_{jk} \tag{3}$$

得到相似度集合为：$S=\{sim(d_i,d_j)\mid i,j=1,2,3,\cdots,19\}$。其中 S 的大小为 $C_{19}^2=171$，为了降低单本图书对聚类结果带来较大的波动，对文档之间的相似性进行归一化处理，则相似度公式(4)如下，图书相似度矩阵如表 5 所示。

$$S_{ij}=\frac{Sim(d_i,d_j)}{\sum S} \tag{4}$$

3.1.3 聚类算法

本文借鉴许厚金等提出的基于相似中心的 k-cmeans 算法来确定初始簇中心，该方法有助于缓解聚类时对初始中心点、“噪声”点、孤立点数据敏感的影响[15]，聚类过程如下：

(1)确定初始簇中心点。采用公式(5)计算每本图书的算子 S_j，按降序排列算子，选择最大值的 k 本图书的空间向量作为初始簇中心，公式(5)中 s_{ij} 表示相似度值。

表 4 《疯狂 Android 讲义》特征词对应权重

特征词	权重 w	特征词	权重 w	特征词	权重 w
讲义	0.1816513	开发	0.2818908	具有	0.1315505
Java	0.2818908	应用	0.2568499	疯狂	0.2568309
Android	0.4071903	编程	0.2317901	全面	0.1315315
本书	0.1865929	阅读	0.2067112	实用	0.1064906
读者	0.2067302	发展	0.1169104	配套	0.1315315
知识	0.1816513	介绍	0.1315505	及时	0.1315315
理论	0.1565914	作为	0.1565914	熟练	0.1315315
界面	0.1565914	处理	0.1315505	合适	0.1315315
图形	0.1565914	提供	0.1565914	认真	0.1315315

表5 资源相似度矩阵

Sij	1	…	5	6	…	12	13	…	18	19
1	1	…	0.020178	0.005101	…	0	0	…	0.006899	0.005696
2	0.01543	…	0.026796	0	…	0.00506	0.017629	…	0.00498	0.009619
3	0	…	0.010912	0	…	0.020873	0.012451	…	0.032416	0.022936
4	0	…	0	0	…	0	0	…	0.008447	0
5	0.020178	…	1	0.003478	…	0.005901	0.003521	…	0.005212	0.004303
6	0.005101	…	0.003478	1	…	0	0.003184	…	0.002794	0.011562
7	0	…	0	0.009418	…	0.005859	0	…	0.007122	0.009295
8	0	…	0	0.003784	…	0	0	…	0	0.004916
9	0.013844	…	0.014316	0.005606	…	0.023704	0.01838	…	0.027903	0.006935
10	0	…	0	0	…	0	0	…	0	0
11	0	…	0	0.003289	…	0	0.019511	…	0.023901	0.009254
12	0	…	0.005901	0	…	1	0.005989	…	0.012808	0.005758
13	0	…	0.003521	0.003184	…	0.005989	1	…	0.009402	0
14	0	…	0	0	…	0	0	…	0.009519	0
15	0.057472	…	0.020563	0.007015	…	0	0.003734	…	0.005526	0.004562
16	0	…	0.007004	0.004143	…	0	0.020557	…	0.025667	0.005383
17	0	…	0.007178	0.003693	…	0	0.014418	…	0	0
18	0.006899	…	0.005212	0.002794	…	0.012808	0.009402	…	1	0.015963
19	0.005696	…	0.004303	0.011562	…	0.005758	0	…	0.015963	1

表6 资源聚类结果

编号	簇群
1	{习惯的力量;围城;我们时代的神经症人格}
2	{心理罪;亲爱的三毛;且听风吟;我是猫;少有人走的路;我承认我不曾历经沧桑}
3	{带一本书去巴黎;失落的玫瑰;美丽新世界}
4	{山南水北;明朝那些事儿;中国近代史}
5	{疯狂 Android 讲义;灵魂机器时代;浪潮之巅;深入理解计算机系统}

$$S_j = \sum_{i=1}^{n} s_{ij}, j = 1, \cdots, n \quad (5)$$

(2)初始簇类。根据前面计算的资源相似度矩阵,将未被选作簇中心的图书分配到相似度最大的簇类下,由此得到 k 个粗分的簇。

(3)更新簇中心。计算各个粗分的簇中所有点的均值,更新为簇中心点,计算方法是取簇中所有图书各维度的算书平均数。

(4)聚类。计算资源空间中所有点与新的 k 个簇中心的相似性,将其分类到相似性最高的簇类中,并计算每个簇的相似性均值,计算方法如公式(6)。

$$C_i = \frac{\sum_{S_j \in S_i} S_j}{|S_i|}, i = 1, \cdots, k; j = 1, \cdots \quad (6)$$

式中:$|S_i|$ 表示第 i 个簇中的图书数量,S_i 表示第 i 个簇中图书集合,C_i 表示第 i 个簇的相似性均值。

(5)停止聚类。如果更新后的簇相似性均值与更新前一致,则停止聚类,否则转步骤(3)。

(6)结果输出。

在以上聚类过程中,k 设置为5,19本图书经过2次迭代后,聚类结果如表6所示。

表7 标签共现矩阵

a_{ij}	IT	计算机	社会	投资	中国	历史	…	心理学
IT		4	0	1	0	1	…	0
计算机	4			1	0	1	…	0
社会	2	2		1	2	2	…	1
投资	1	1	1		0	1	…	0
中国	0	0	2	0		2	…	1
历史	1	1	2	1	2		…	0
…	…	…	…	…	…	…		…
心理学	0	0	1	0	1	0	…	

表8 共现标签频度

标签	IT	计算机	社会	投资	中国	历史	…	心理学
$\Gamma(ti)$	29	29	39	9	37	34	…	30

3.2 标签聚类

采用标签共现分析来实现标签聚类，如果两个标签共同标注的资源越多，则两者相似度越高[14]。在王娅丹等提出的基于标签共现的标签聚类算法[16]之上加以调整，完成标签聚类（热门标签）工作。

3.2.1 向量表示

通过计算标签共现频度，及标签在每篇图书中的重要性，获得标签在整个资源集中的重要度矩阵，以此作为标签的特征向量。具体计算过程如下：

(1)计算标签共现矩阵 A_{n*n}。n 为标签总数，矩阵 $A_{n\times n}$ 是 $n\times n$ 型矩阵，则 a_{ij} 表示标签 t_i 和标签 t_j 标注过同一资源的次数，如表7所示：

该矩阵表示，在一定程度上，a_{ij} 越大说明标签 t_i 和 t_j 共同出现的几率越高，即标签 t_i 和标签 t_j 之间的关系就越强。

(2)共现标签频度向量 $\Gamma(t_i)$。$\Gamma(t_i)$ 表示与标签 t_i 在同一资源中的同时出现过的标签的个数，如表7所示：

(3)标签重要度矩阵 $L_{n\times n}$。n 表示标签数量，该矩阵中的元素 l_{ij} 表示标签 t_i 在资源集中的重要程度，计算公式(7)如下：

$$l_{ij} = a_{ij} \times \lg\left(\frac{n}{1+\Gamma(t_i)}\right) \tag{7}$$

在公式(7)中，为了平滑分母，使分母加1，消除共现频度为0的标签的影响。结果如表9所示。

该矩阵表示在资源集合中，出现频率高、且与其他标签共现频率低的标签更重要，即 l_{ij} 越大，表示标签 t_i 在资源集合中越重要。每个标签所在的行向量即代表该标签的特征向量。

3.2.2 相似度计算

与计算资源相似度方法类似，同样使用余弦相似度计算两个向量间的相似性，得到标签相似度矩阵，计算公式(8)如下：

$$S_{ij} = Sim(L_i, L_j) = \frac{\sum_{k=1}^{n} l_{ik} l_{jk}}{\sqrt{\sum_{k=1}^{n} l_{ik}^2 \sum_{k=1}^{n} l_{jk}^2}} \tag{8}$$

式中，L_i 表示标签 i 在矩阵 $L_{n\times n}$ 中的行向量，即标签 i 的特征向量 $L_i(l_{i1}, l_{i2}, \cdots, l_{ik}, \cdots, l_{in})$，公式(8)反应了两两标签间线性相关程度的统计量。计算结果见表10。

3.2.3 聚类算法

采用与资源聚类同样的方法进行聚类，经过4次迭代得到结果如表11。

3.3 用户聚类

大众标注包含三个基本元素：资源、用户和标签（指用户标签）。可以用四元组 D 来表示：$D = <U, R, T, A>$，U 是用户集合，R 是资源集合，T 是标签集合，A 是标注关系集合。A 可以表示为：$A \subseteq \{<u, r, t>: u \in U, r \in R, t \in T\}$。

表9 标签重要度矩阵

l_{ij}	IT	计算机	社会	投资	中国	历史	…	心理学
IT		2.319134	0.90969	1.056905	0	0.512837	…	0
计算机	2.319134		0.90969	1.056905	0	0.512837	…	0
社会	0.90969	0.90969		1.056905	0.954243	1.025674	…	0.56554
投资	1.056905	1.056905	1.056905		0	0.512837	…	0
中国	0	0	0.954243	0		1.025674	…	0.56554
历史	0.512837	0.512837	1.025674	0.512837	1.025674		…	0
…	…	…	…	…	…	…		…
心理学	0	0	0.565543	0	0.565543	0	…	

表10 标签相似度矩阵

S_{ij}	IT	计算机	社会	投资	中国	历史	…	心理学
IT		1.97196	0.83966	0.92805	0.32576	0.96258	…	0.1806
计算机	1.97196		0.62213	0.62839	0.29607	0.61433	…	0.1806
社会	0.83966	0.62213		0.78124	0.68787	0.48854	…	0.6021
投资	0.92805	0.62839	0.78124		0.19034	0.60389	…	0.2018
中国	0.32576	0.29607	0.68787	0.19034		0.98643	…	0.9415
历史	0.96258	0.61433	0.48854	0.60389	0.98643		…	0.2968
…	…	…	…	…	…	…		…
心理学	0.18058	0.18058	0.60208	0.20178	0.94151	0.29677	…	

表11 标签聚类结果

编号	簇群
1	{Android;Google;IT;计算机;社会;投资;编程;程序员;互联网;机器;计算机科学;技术;讲义;教材;开发;科技;科普;科学;人工智能;软件;商业;设计;实用;未来;系统;云计算}
2	{纪实;明史;当年;感触;感悟;近代;近代史;经典;历史;明朝;明月;社会学;徐中约;有趣}
3	{思想;新世界;村上春树;讽刺;孤独;赫胥黎;荒诞;觉醒;科幻;林少华;猫;名著;且听风吟;日本;神秘;土耳其;外国;乌托邦;夏目漱石;小说;英国;政治}
4	{罪;管理;励志;思维;爱;犯罪;精神分析;卡伦霍妮;苦难;雷米;人性;神经病;推理;完善;力量;效率;心理;心理学;心灵;心智;悬疑;哲学}
5	{爱情;法国;旅游;人生;闲书;现代;游记;杂文;中国;巴黎;成长;当代;韩少功;婚姻;蒋方舟;林达;旅行;钱钟书;青春;三毛;散文;生活;随笔;台湾;围城;文化;文学;杂文}

建立用户在资源集、标签集、标签－资源关系集上的行向量，分别表示为 $R_j=\{x_i|x_i\geqslant 0\}$、$T_j=\{y_i|y_i\in 0,1\}$、$A_j=\{z_i|z_i\in 0,1\}$，其中 x_i 表示如果用户拥有资源 i，则 $x_i=1$，否则 $x_i=0$，y_i 表示用户使用标签 i 的频数，z_i 表示如果用户使用标签标注过资源 i，则 $z_i=1$，否则 $z_i=0$。R_j 的大小为资源数量 M，T_j 的大小为标签数量 N，A_j 的大小为 $M\times N$。例如用户 Z.Y.T 关于资源和标签的行向量为：$R_1=\{1,1,1,1,$

$1,1,0,0,0,0,0,0,0,0,0,0,0,0\}$，$T_1=\{2,\underbrace{1,\cdots,1}_{11},2,\underbrace{1,\cdots,1}_{9},\underbrace{0,\cdots,0}_{42}\}$。

定义用户 A 与用户 B 的相似度计算公式(9)如下：

$$S_{1,2}=\alpha w_r+\beta w_t+\gamma w_a \tag{9}$$

其中，w_r表示用户 U_1和用户 U_2在资源集上的相似度权重，w_t表示在标签集上的相似度权重，w_a表示在标签－资源关系集上的相似度权重(w_r、w_t、w_a都经过归一化处理)，a、β、y 分别为对应权重的系数，结果如表 12、表 13、表 14 所示。

表 12　资源集上的相似度

w_r	Z. Y. T	泥蓝海馨	Andrea	徐徐又当这信
Z. Y. T	1	0.28	0.52	0.32
泥蓝海馨	0.28	1	0.36	0.32
Andrea	0.52	0.36	1	0.32
徐徐又当这信	0.32	0.32	0.32	1

表 13　标签集上的相似度

w_t	Z. Y. T	泥蓝海馨	Andrea	徐徐又当这信
Z. Y. T	1	0.1667	0.1667	0.1667
泥蓝海馨	0.1667	1	0.1667	0.5
Andrea	0.1667	0.1667	1	0
徐徐又当这信	0.1667	0.5	0	1

表 14　标签－资源关系集上的相似度权重

w_a	Z. Y. T	泥蓝海馨	Andrea	徐徐又当这信
Z. Y. T	1	0.14286	0.14286	0.21429
泥蓝海馨	0.14286	1	0.14286	0.35714
Andrea	0.14286	0.14286	1	0
徐徐又当这信	0.21429	0.35714	0	1

为了方法论述清晰，认为资源和标签重要程度一致，采用同一标签标注同一资源的重要程度认定比前两种情况更大，因此设置 a 和 β 为 0.25，y 为 0.5。最后得到用户相似度如表 15 所示。虽然本研究选用的用户量很少，但是为了研究方法的论述完整性，此处仍对用户加以聚类，仍然采用与前面一直的聚类方法，结果如表 16 所示。

表 15　用户相似度矩阵

S_{ij}	Z. Y. T	泥蓝海馨	Andrea	徐徐又当这信
Z. Y. T	1	0.1831	0.2431	0.22894
泥蓝海馨	0.1831	1	0.2031	0.3836
Andrea	0.2431	0.2031	1	0.08
徐徐又当这信	0.2289	0.3836	0.08	1

表 16　用户聚类结果

编号	簇群
1	{泥蓝海馨;Andrea}
2	{徐徐又当这信;Z. Y. T}

3.4　聚类结果分析

观察表 6、表 11、表 16 所分别对应的资源、标签、用户聚类结果发现，在不考虑用户、资源、标签之间的关系而对各项分别进行聚类，结果并不准确。表 6 是根据资源内容特征的相似度进行聚类的，显然簇群 1、簇群 2、簇群 3 较为混乱，比如《围城》和《亲爱的三毛》、《且听风吟》等同属小说，却没有聚类到一起，《习惯的力量》和《少有人走的路》同属心理学书籍也被分到不同簇群下，如果直接依据资源的聚类结果进行推荐，则只能推荐《山南水北》和《中国近代史》，同样为历史类书籍的《带一本书去巴黎》便会被遗漏。表 11 是依据标签在资源中的共现频次及重要度计算相似性后得到的聚类结果，在聚簇过程中，发现与 5 个簇群的中心距离最近的标签分别是：计算机和 IT、历史、小说、心理、中国，而对图书分析后可用五种标签分类：计算机、历史、小说、心理、散文，前后稍有误差，这表明，标签聚类的结果也不是完全准确，如“三毛”属于簇群 2 小说类中被分到了簇群 5 中，并且标签“小说”也有被用于标注历史书籍《明朝那些事儿》，所以单纯依赖于标签实现资源的分类，也是不合理的。用户的聚类结果表 16，是依据用户间标签、资源的联系聚类得到的，并未考虑热门标签对检索的影响。

4　基于组合条件的个性化推荐模型构建

基于 3.4 对聚类结果的分析，构建组合推荐模型以消除分别聚类带来的结果不准确。图 1 呈现了推荐模型的总体框架，这一部分将分析资源重组的

3个组件:基于相似资源的推荐组件、基于用户检索词的推荐组件和基于相似用户的推荐组件。

4.1 推荐组件描述

(1)基于相似资源的推荐组件

该组件是以资源间相似度作为推荐算法的核心。计算网站中所有资源间的相似度并保存,对资源依据相似度进行聚类分簇,获取目标资源与所有簇群中相似度最大的资源集合作为该组件下的推荐资源集。

算法1:训练已有资源集合 $R=\{r_1,r_2,\cdots,r_p\}$,每条资源用特征向量表示为 $r=\{d_1,d_2,\cdots,d_M\}$,使用调整后的K-means算法对资源集合 R 进行聚类成簇。测试目标资源 $r^*\in R$,则 r^* 所在簇群集合为 R^*,选择与 r^* 相似度最大的 K_1 本图书作为候选推荐集 Book_1。

根据算法1,得到候选推荐集 $Book_1${中国近代史;山南水北},对相似度进行归一化,结果如表17。

表17 候选推荐资源集 Book_1

相似度	中国近代史	山南水北
明朝那些事儿	0.0691	0.0504

(2)基于用户检索词的推荐组件

匹配用户检索词与网站已有标签(指热门标签),是检索最为快速的一种方式。对检索语句进行特征词处理,表示为空间特征向量形式。

算法2:考虑已有的资源集合 $R=\{r_1,r_2,\cdots,r_N\}$,及其对应的标签集合 $T=\{t_1,t_2,\cdots,t_M\}$,采用调整后的K-means算法对标签集合 T 进行聚类成簇。测试检索术语 t^*,将其自动分类到最近的簇中心下,选择该簇群下与 t^* 最为相似的 K_2 个标签构成相似标签集合,该集合中标签标注过的资源,以标注重合度高低排序,选择靠前的 K_2 本作为候选推荐集 Book_2。

选用类中心向量距离法作为自动分类算法。根据算法2,得到相似标签集合{明史;当年;感触;感悟;经典;历史;明月有趣}及其权重如表18,依据其标注过的资源得到候选推荐资源集 Book_2,见表19。

(3)基于相似用户的推荐组件

一般认为"相似"(距离近)的人有相似的兴趣与偏好,相似的资源被相似的用户所喜爱,所以这一部分将通过寻找与发出检索要求的用户相似的用户来进行资源推荐。

算法3:考虑已有用户集 $U=\{u_1,u_2,\cdots,u_p\}$,用户对应的资源集合为 $R=\{r_1,r_2,\cdots,r_N\}$,用户对资源标注的标签集合为 $T=\{t_1,t_2,\cdots,t_M\}$,P、N、M 分别表示用户及其相关资源、标签的数量,采用简化后的基于超图投影和节点相似性算法对用户集 U 聚类成簇。测试目标用户 $u^*\in U$,选择 u^* 所在簇群下最相近的 K_3' 个用户构成相似用户集合,该集合中用户拥有的资源,按拥有重合度高低排序,选择靠前的 K_3 本作为候选推荐集 Book_3。

根据算法3得到相似用户"Andrea",将该用户拥,的资源构成候选推荐资源集 Book_3,结果如表19。

表18 相似标签集合

相似度	明史	当年	感触	感悟	经典	历史	明月	有趣
明朝	8.9364	8.9364	10.0534	8.9364	4.6445	4.8782	10.0534	8.9364

表19 候选推荐资源集 Book_2

书籍	我承认我不曾历经沧桑	带一本书去巴黎	浪潮之巅	围城	中国近代史	我们时代的神经症人格	深入理解计算机系统
相似度	0.2619	0.1271	0.1271	0.1210	0.1210	0.1210	0.1210

表20 候选推荐资源集 Book_3

书籍	我们时代的神经症人格	美丽新世界	且听风吟	中国近代史	浪潮之巅	深入理解计算机系统
相似度	0.0177	0.0283	0.009	0.2077	0.0210	0.0120

表 21　个性化推荐资源集 $Book_4$

书籍	我们时代的神经症人格	浪潮之巅	深入理解计算机系统	我承认我不曾历经沧桑	且听风吟	美丽新世界	带一本书去巴黎
相似度	0.3803	0.2884	0.2688	0.2619	0.2478	0.1838	0.1271

4.2　资源重组

由 3 种组件下推荐的候资源集进行重组得到最后的个性化推荐资源。

算法 4：对于任意目标用户 u^*，将前三个组件推荐的候选资源集归一化后的权重相加，把所有 u^* 没有选择过的产品按照权重值进行排序，并把排名靠前的资源推荐给 u^*。

根据 4.1 中的算法 1、算法 2、算法 3 分别得到各组件下的推荐资源集，并对候选推荐资源集中资源的权重进行归一化处理，由算法 4 对三个候选推荐资源集重组后，得到最后的推荐资源集合，如表 20 所示，即在本模型下，根据用户“泥蓝海馨”在检索图书《明朝那些事儿》后，获得了该用户还可能感兴趣的图书 $Book_4$。

4.3　推荐结果分析

表 17、表 19、表 20 分别是在 3 种组件下得到的推荐资源集，可以发现单纯基于资源的推荐，推荐结果较精确，但过于注重内容相似度却造成结果不完整，也不具有新颖性；单纯基于标签的推荐，结果准确度不高，资源类型多而不一，比如《深入理解计算机系统》、《围城》、《我们时代的神经症人格》与《明朝那些事儿》的相关性没有统一判断标准；而基于用户的推荐可以使资源推荐不再局限于资源的内容相似性，这可以为用户带来意外惊喜，但也忽视了资源间的内容相关性。这三种方法正是以往国内外学者研究社会化标注系统中个性化信息推荐的着眼点，任何单一的推荐方法都有其局限性。基于资源相似度的推荐结果会使得资源重叠度高、扩充性不够，对用户的有用性较低；基于标签的资源推荐，由于用户标签质量不一、形式多样，仅以此为基础研究个性化信息推荐存在与生俱有的缺陷[17]；基于邻近用户的信息推荐方法由于用户特征维度难以确定，且偏好和情感态度等在变化之中，目前对此研究较为薄弱，不适合用作唯一推荐基础。因此，移植、借鉴、融合并优化已有方法采取组合方式推进社会化标注系统中信息资源推荐的个性化应是未来研究的重点[18]，本文将三个组件计算得出的候选资源集合进行重组，充分考虑了基于内容与协调过滤推荐技术的优缺点及社会化标注系统的特点，从资源、标签、用户三个维度结合推荐，使用标签和资源表征用户，这对用户兴趣偏好、情感态度的表达更客观且动态性更强，而揭示资源间、标签间相似性、关联性是难点也是重点。本文提出的方法结果如表 21 所示，取得了较好的推荐效果，这对未来提高个性化推荐的准确度、避免推荐技术中的冷启动、数据稀疏等困难提出了新的思路。

5　结　语

以标签和用户作为个性化推荐研究的对象已经得到了国内外大多学者的关注，但是以资源作为目标推荐对象的并不多，且在以资源作为推荐对象的研究中，研究者主要单一的采用一种推荐方法[19]，或是基于“用户 - 标签 - 资源”的资源聚合与推荐方法，而研究三元组中各部件分别与资源的关系，应用组合思想进行资源的个性化推荐并未受到学者关注。本文结合已有推荐模型的优势，提出了一种资源重组的方法，构建推荐系统模型，实现在社会化标注网站中为用户个性化推荐资源。笔者在对以往关于社会化标注系统及推荐技术的学习研究，并使用了国内的部分社会化标注网站，进行信息检索及对个性化推荐结果的效用体验后，提出了推荐系统的总体框架，包含三个部件：基于相关资源的推荐组件、基于相关标签的推荐组件、基于邻近用户的推荐组件。在每个组件中选用合适有效的相似度计算方法和聚类算法实现组件的资源推荐功能，最后，对 3 个组件加以重组获得个性化推荐资源集。从实证结果来看，提出的模型个性化推荐效果理想，但为了模型描述清晰未选用大数据量，今后的研究方向将扩大实例数据的规模，完善模型中的相关算法，实现提出的模型系统，同时用实证结果对模型进行评估，以便更好地验证算法模型的可行性和有效性，促使算法模型从理论走向实践。

参 考 文 献

[1] Wikipedia. Thomas Vander Wal. Folksonomy. [EB/OL]. (2013-04-27) [2015-12-05] https://en. wikipedia. org/wiki/Thomas_Vander_Wal#Folksonomy

[2] Hotho A, Jäschke R, Schmitz C, et al. Information Retrieval in Folksonomies: Search and Ranking [J]. Semantic Web Research & Applications, 2006, 4011: 411-426.

[3] Surabhi D, Thorat C. Namrata Mahender. SOCIAL TAGGING IN SOCIAL MEDIA - A REVIEW. [J]. International Journal of Research in Computer Applications & Robotics, 2015, 3(3):130-137.

[4] 程慧荣，黄国彬，张永杰. 国外大众标注系统研究进展[J]. 图书馆杂志，2008(11):54-59.

[5] Daniela G, Alejro C. Folksonomy-Based Recommender Systems: A State-of-the-Art Review [J]. International Journal of Intelligent Systems, 2015.

[6] 刘建国，周涛，汪秉宏. 个性化推荐系统的研究进展[J]. 自然科学进展，2009，19(1):1-15.

[7] Arekar T, Sonar M R S, Uke N J. A Survey on Recommendation System [J]. International Journal of Innovative Research in Advanced Engineering, ISSN: 2349-2163, Volume 2 Issue, January 2015.

[8] Adomavicius G, Tuzhilin A. Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions [J]. IEEE Transactions on Knowledge & Data Engineering, 2005, 17(6):734-749.

[9] 陈祖琴，葛继科. Web2.0环境中基于社会标注的个性化推荐系统模型研究[J]. 电子商务，2012(2): 60-62.

[10] 张玉. 基于社会化标签的个性化推荐系统研究[D]. 合肥:合肥工业大学，2011.

[11] 吴思竹. 社会标注系统中标签推荐方法研究进展[J]. 图书馆杂志，2010(3):48-52.

[12] 张林东. 一颗长势良好的"豆瓣"[J]. 上海信息化，2007(5):76-79.

[13] Peter Harrington. 机器学习实战[M]. 李锐等译. 北京：人民邮电出版社，2013:184-185.

[14] 熊回香. 面向Web3.0的大众分类研究[D]. 武汉：华中师范大学，2011.

[15] 许厚金，刘永炎，邓成玉,等. 基于相似中心的k-cmeans文本聚类算法[J]. 计算机工程与设计，2010，31(8):1802-1805.

[16] 王娅丹，李鹏，金瑜,等. 标签共现的标签聚类算法研究[J]. 计算机工程与应用，2015(2):146-150.

[17] 易明，邓卫华. 基于标签的个性化信息推荐研究综述[J]. 情报理论与实践，2011，34(3):126-128.

[18] 毕强，赵夷平，孙中秋. 社会化标注系统资源聚合的实证分析[J]. 情报资料工作，2015(5):30-37.

[19] 易明，邓卫华，徐佳. 社会化标签系统中基于组合策略的个性化知识推荐研究[J]. 情报科学，2011(7): 1093-1097.

（责任编辑 王海燕）