一种基于动机倾向的标签推荐方法*

靳延安
(湖北经济学院 信息管理学院, 武汉 430205)

摘　要: 为了能够推荐符合用户信息需求的标签,在深入分析社会标签空间和传统标签推荐方法的基础上,提出了度量用户和资源的动机倾向性的五种指标,并对其测度有效性进行了验证。基于此指标体系,建立了动机倾向性判别模型,并设计了推荐算法。实验结果表明,基于动机倾向的推荐算法比当前主流推荐算法具有更加准确的推荐结果。

关键词: 社会标签; 标注; 推荐系统; 用户动机
中图分类号: TP391　　**文献标志码**: A　　**文章编号**: 1001-3695(2013)01-072-06
doi:10.3969/j.issn.1001-3695.2013.01.017

Approach for tag recommendation based on orientation of motivation

JIN Yan-an
(*Schoole of Information Management*, *Hubei University of Economics*, *Wuhan* 430205, *China*)

Abstract: In order to recommend tags meeting the demand of social taggers, after deeply analyzing social tag spaces and previous recommendation methods, this paper used five indices to measure the users' and resources' motivation. After validating the five metric indices, it proposed an orientation of motivation discrimination model (OMDM), and designed an algorithm to recommend tags based on the model. Experimental results show, that the proposed method can provide more accurate candidate tags than current mainstream methods.

Key words: social tags; tagging; recommendation system; user motivation

0　引言

社会标签是 Web 2.0 时代的一个伟大创举,它将网络信息资源的分类方式从原来基于少数专家的分类体系转变为广大 Web 用户的分类体系。Web 用户可以使用任意的词汇不受控地对网络信息资源进行分类,利用社会标签系统定义自己的分类,同时这种分类还可以被其他用户所分享。

但是社会标签系统中标签不受控制地使用,给基于标签的分类系统带来了许多问题。出于对信息和词的不同理解,不同的用户不太可能使用完全一致的方法分类相同或者相似的信息。研究结果表明,基于标签的分类系统通常难以保证分类的一致性,并面临着冗余性、不完备性等问题。这些问题导致基于标签的分类系统在信息检索等领域中的实际应用效果大打折扣。

为了保证分类的一致、完备,研究人员对标签推荐方法进行了大量的研究[1]。标签推荐的基本方法是通过为用户在进行基于标签的分类过程中提供高质量的标签备选。截至目前,研究人员已经提出了大量的标签推荐方法[2~4],主要有基于网络结构的方法、基于张量的方法和基于主题的方法。这些方法都能从一定程度上解决其中的一个或几个问题,但依然存在冷启动、稀疏性和多样性等问题,最终导致用户不会从所推荐的备选列表中选用标签。笔者认为,如果能够捕捉到标注用户的标注动机,弄清楚用户为什么要标注 Web 信息资源及其标注习惯,据此进行推荐的效果会好于传统方法。因此,本文研究用户使用标签的一般规律,试图从社会标签用户进行标注的动机出发为用户推荐社会标签。

动机是推动人从事某种活动并朝一个方向前进的内部动力,是为实现一定目的而行动的原因。心理学认为,动机是个体的内在过程,行为是这种内在过程的表现[5]。因此,人们不能直接观察动机,但可以间接推断。例如,用户标注"搜狐主页"是为了便于浏览具体的网页内容、还是为了便于查找该网站,他的目标是什么;用户在进行标注时,是一直喜欢用中性词进行标注、还是喜欢用有感情的词进行标注等,可以通过诸如此类用户标注的行为和结果来间接研究用户标注的动机。

表 1 汇总了有关社会标签系统中用户标注行为动机的研究。从表 1 可以看出,从理论证明[2]、证据观察[6,7]到大规模数据的实证研究[8~11],从早期的定性研究到后来的定量研究,关于标注行为动机的分类、评价和研究范围没有一致性的意见。Sinha[12]对标注的整个过程进行了分析,认为标注过程分为有关资源的概念产生和概念选择两个阶段。文献[4,6]将用户动机绝对二分为分类动机和描述动机,即要么是描述动机,要么是分类动机。显然,该分类方法并不符合现实情况。更客观地,应该是说用户更倾向于哪一种动机。倾向于描述的用户用标签来概括资源的内容,这些标签可以方便用户检索到资源;而倾向于分类的用户把标签看做分类的类目,这些标签可以帮助用户管理和浏览资源。

本文将用户标注的动机分为描述倾向(动机)和分类倾向(动机)。通过深入分析众多社会标签系统和传统标签推荐方

收稿日期: 2012-06-13; **修回日期**: 2012-07-16　　**基金项目**: 国家自然科学基金资助项目(70771043)
作者简介: 靳延安(1975-),男,河南郑州人,讲师,博士,CCF 会员,主要研究方向为 Web 智能处理、数据挖掘(yan.an.jin@hbue.edu.cn).

法,提出用户动机倾向可以通过五种指标进行测度,并对其测度有效性进行了验证。在此基础上,建立了可以判别用户动机倾向的判别模型(orientation of motivation discrimination model, OMDM)。根据此模型,设计了基于动机倾向性的标签推荐算法。

表 1　社会标注用户动机研究一览

作者	动机分类
Hotho 等人[2]	what it is about, who owns it, what it is, refining categories, self reference, identifying qualities, task organizing
Tom[6]	description, categorization
Hammond 等人[7]	self/self, others/self, self/others, others/others
Heckner 等人[8]	resource sharing, personal information management
Wash 等人[9]	sharing, later retrieval, social recognition, others
Ames 等人[10]	social/communication, social/organization, self/organization, self/communication
Nov 等人[11]	enjoyment, reputation, self development, commitment
Marlow 等人[13]	future retrieval, attract attention, self presentation, play and competition, contribution and sharing, opinion expression
Xu 等人[14]	subjective, organization, content-based, attribute-based, context-based
Sen 等人[15]	Self-expression, decision support, organizing, learning, finding
Strohmaier 等人[16]	Description, categorization

首先找到与待标注资源有相似动机倾向性的其他资源;然后筛选出与用户具有相似动机倾向性的资源,并聚合它们的标签;将这些标签作为候选推荐对象,依次计算候选推荐对象和待标注资源内容的相关性,将相关性大的标签推荐给用户。实验结果表明,该算法比当前主流推荐算法具有更加准确的推荐结果。

1　动机倾向性的类型及度量

1.1　动机倾向性的类型

由于用户的标注动机不能严格绝对地区分,本文参照文献[4,6]的分类方法,将标注动机倾向性分为描述倾向和分类倾向。

1.1.1　分类倾向

为了便于浏览和管理资源,分类倾向的用户在语义启动过程[12]的第二个阶段,会选择没有相同含义而又概念明确、客观、区分度强的标签。用户长期使用标签系统之后,就会形成一个没有冗余的简单而又稳定的词表,并会限制词表规模。例如,具有分类倾向用户在标注一个汽车图片时,可能会使用 car,而不会使用 automobile、vehicle 等具有相同含义的概念。

1.1.2　描述倾向

为了便于查询与检索资源,描述倾向的用户会从不同角度选择概念作为标签。因此,在语义启动过程第二阶段,用户不仅会选择第一阶段产生的所有概念作为标签,而且会联想更多的概念。例如,描述倾向的用户会使用 car、vehicle、Toyota、wheel 来描述一个汽车图片。所以,描述倾向用户的词表具有规模大、开放性强、变化多端的特点。

1.2　倾向性的度量指标

1.2.1　符号定义

为了建立基于动机倾向性的推荐模型,定义以下符号来表示社会标签系统中的各元素:

U 为用户的集合,即 $U=\{u_1,u_2,\cdots,u_n\}$,u_i 为某一个用户;

T 为标签的集合,即 $T=\{t_1,t_2,\cdots,t_n\}$,t_i 为某一个标签;

R 为资源的集合,即 $R=\{r_1,r_2,\cdots,r_n\}$,r_i 为某一个资源,可能是一个图片,也可能是一个网页;

T_u 为用户 u 使用过的所有标签,即用户 u 的词表;

$|T_u|$ 为用户 u 的词表规模;

T_r 为资源 r 被赋予的所有标签;

R_u 为用户 u 标注过的所有资源;

$|R_u|$ 为用户 u 标注过的资源数;

$R_u(t)$ 为用户 u 使用标签 t 标注过的资源。

1.2.2　资源的标签使用率

资源的标签使用率 TRR_u(tag/resource ratio)即用户 u 为每个资源赋予标签的平均个数,它等于 $|T_u|/|R_u|$。为了方便建立判别模型,进行归一化处理,如式(1)所示。

$$\mathrm{TRR}_u=\mathrm{e}^{-|T_u|/|R_u|} \tag{1}$$

为了保证分类的一致性和应用方便,用户分类倾向会选择具有区分度而又无歧义的少数代表性词汇作为标签。而描述倾向的用户则刚好相反,为了更全面地揭示资源,往往会从不同角度或不同层面选择各种各样的词来描述资源。

显然,$\mathrm{TRR}_u\in(0,1)$。一般来讲,$\mathrm{TRR}_u\to1$,用户 u 越倾向于描述;$\mathrm{TRR}_u\to0$,则该用户 u 越倾向于分类。

1.2.3　低频标签使用率

用户在标注少数特别资源时才可能会用到的标签称之为低频标签。用户 u 的低频标签使用率 LFTR_u(lower frequency tag ratio)即低频标签的数量占用户 u 标签总数的比例,记为 LFTR_u,其度量用式(2)来计算。

$$\mathrm{LFTR}_u=|T_u^{\circ}|/|T_u|,T_u^{\circ}=\{t\,|\,|R(t)|\leqslant n\},n=\lceil|R(t_{\max})|/100\rceil \tag{2}$$

其中:$|T_u^{\circ}|$ 为低频标签的数量,n 为最频繁使用标签次数的百分之一。

显然,$\mathrm{LFTR}_u\in(0,1)$。当 $\mathrm{LFTR}_u\to1$ 时,意味着用户 u 使用了较多的低频词,用户使用这些低频词的目的就是为了全面揭示资源的方方面面,用户 u 不在乎低频词的使用带来的其他问题,此时,可以认为用户 u 具有明显的描述倾向;当 $\mathrm{LFTR}_u\to0$ 时,意味着用户 u 几乎不使用低频标签,这类用户认为低频标签等同于噪声,势必造成分类类目多样性,甚至使分类不一致,进而不利于浏览,此时,可以认为用户 u 具有明显的分类倾向。

1.2.4　标签语义重复因子

分类倾向的用户为了确保分类的一致性,在选用标签时会力求避免选用语义相同的词。但为了将来能够有更高的召回率,描述倾向的用户希望通过标签来全面描述资源,因而可能会使用多个语义重复的标签。因此,可通过考察用户词表中标签的语义相似度来衡量用户所具有的动机倾向。本文使用 TSOF_u(tag semantic overlap facto)来度量用户 u 使用语义重复的标签情况,如式(3)所示。

$$\mathrm{TSOF}_u=2\sum\mathrm{sim}(t_i,t_j)/|T_u|(|T_u|+1)\quad i\neq j \tag{3}$$

其中:$\mathrm{sim}(t_i,t_j)$ 表示标签 t_i、t_j 的相似性,使用文献[17]中式(13)来计算 $\mathrm{sim}(t_i,t_j)$。当用户 u 的词表中所有标签语义上都

相同时,$TSOF_u \to 1$,说明用户 u 有描述倾向;反之,说明用户 u 有分类倾向。

1.2.5 标签的相对条件熵

从信息论的角度看,用户选择标签的过程就是为资源进行编码的过程。对于分类倾向的用户来说,他们希望标签有最大的区分度,实际上就是希望为特定信息的编码最短,即条件信息熵最大。根据信息熵的知识,每个标签的使用率相同时,标签的条件信息熵最大。也就是说,所有标签都有均等机会被使用到。而对于描述倾向的用户来说,他们并不关心标签的区分度。可以按照式(4)来计算标签的条件信息熵。

$$H_u(R|T) = -\sum_{r\in R}\sum_{t\in T} p(r,t)\log_2 p(r|t) \tag{4}$$

其中:$p(r,t)$为标签在资源上的分布。为了能够区分用户之间的差别,对条件熵进行归一化处理以保留编码信息,用实测的条件熵 $H_u(R|T)$和理想的条件熵 $H_{opt}(R|T)$进行比较。理想的条件熵 $H_{opt}(R|T)$在每个标签都有相同使用率的情况取得。在此基础上,用户 u 标签的相对条件熵 $TRCE_u$(tag relative conditional entropy)计算如式(5)所示。

$$TRCE_u = [H_{opt}(R|T) - H_u(R|T)]/H_{opt}(R|T) \tag{5}$$

显然,$TRCE_u \in (0,1)$。当用户 u 把标签用做分类时,标签的区分能力最强,条件熵也最接近理想情况,此时,$TRCE_u$ 也越接近 0,可以认为用户也倾向于分类;反之,可以认为用户倾向于描述。

1.2.6 疑问副词标签使用率

一般情况下,用户不会使用 when、how、what 等疑问副词作为标签。但笔者对实验所用数据进行处理时发现,疑问副词作为标签时,该次标注的其他标签往往是资源标题中的实词。对使用疑问副词作为标签的这类用户进行分析,发现这类用户具有明显的描述倾向,如图 1 中用户 breneaux 的标注记录。

图1 用户breneaux的标注记录

因此,疑问副词作为特殊标签的使用率可以作为用户标注倾向性的判别指标之一。如果用户使用疑问副词标签的比例很高,那么该用户可判定为具有描述倾向;反之,可判定为具有分类倾向。疑问副词标签使用率 STR_u(special tag ratio)计算如式(6)所示。

$$STR_u = \mathrm{card}(t \in T_{str})/|T_u| \tag{6}$$

其中:$T_{str} = \{\text{what}, \text{who}, \text{when}, \text{where}, \cdots\}$;$\mathrm{card}(t \in T_{str})$为用户 u 使用疑问副词作为标签的个数,含重复计数。显然,$STR_u \in (0,1)$。当 $STR_u \to 1$ 时,用户 u 越可能具有描述倾向;当 $STR_u \to 0$ 时,用户 u 越不可能具有描述倾向。

1.3 度量指标间相关性检验

为了更好地建立推荐模型,本文采用 Spearman 系数(表 2)对用户标注倾向性的度量指标间的相关性进行检验。检验所使用的样本数据为推荐实验所用的两个数据集,每个数据集包含 100 个用户所有资源的所有标注。每种度量指标都是从不同的方面来度量用户的标注动机。但从表 2 的检验结果来看,相关性较高的是 TRR 和 TSOF,这两个指标本质上是一致的,其他的三个指标中 LFTR 和 TRCE 高度相关。相对来说,五个指标中 STR 和其他指标的相关性要小些。这些高度相关的指标预示着可能有相似的动机。例如,为提高浏览效率和分类一致性,分类倾向的用户会尽可能使用少而不重复的标签,这样他们的 TRR 和 TSOF 就会比较低。同时,他们希望尽可能少地使用低频标签,尽可能使得每个标签都有相同的使用率,即使得 LFTR 和 TRCE 尽可能的低。

表 2 五种动机倾向性指标的 Spearman 系数

指标	TRR	LFTR	TRCE	TSOF	STR
LFTR	0.83				
TRCE	0.81	0.89			
TSOF	0.92	0.83	0.79		
STR	0.52	0.56	0.49	0.59	

2 用户倾向性判别

2.1 用户倾向性判别分析

社会标签系统中并不要求用户具体描述使用标签的目的和解释选择某一标签的原因。但在标注时,用户实际上在短时间内启动了一个语义建立过程,即当用户浏览资源时,用户在脑海中就形成了一些与资源相关的概念,然后从众多概念中选择合适的概念赋予资源[12,18]。虽然这个过程瞬间完成,但一定程度上还是能反映一段时间内用户相对稳定的信息需求。

虽然不能直接度量标注用户的动机,但可以利用标注的结果进行反推。本文使用了五种指标对用户所使用的标签历史进行度量。基于这种度量,构建判别用户动机倾向性的模型如下:

$$M_u = c_1 * TRR + c_2 * LFTU + c_3 * TRCE + c_4 * TSOF + c_5 * STR \tag{7}$$

其中:$c = (c_1, c_2, c_3, c_4, c_5)$,$c_1, c_2, c_3, c_4, c_5 \in (0,1)$为判别系数;TRR、LFTU、TRCE、TSOF、STR 为 1.2 节中所构建的度量指标,这些指标的取值均在 0 ~ 1 间。显而易见,$M_u \in (0,1)$。根据 1.2 节中每个指标的含义,很容易推理出以下结论:

a) M_u 具有单调性。

b) 当 $M_u < M_{threshold}$ 时,用户 u 的动机趋于分类;当 $M_u > M_{threshold}$ 时,用户 u 的动机趋于描述,$M_{threshold}$ 为分类动机和描述动机的临界值。

因此,在判定用户动机倾向性时,首先计算该用户的五个度量指标,并将计算结果代入推荐模型计算得到 M_u,将 M_u 和临界值 $M_{threshold}$ 进行比较来确定用户的动机倾向性,按照式(8)进行判别。

$$\begin{cases} M_u > M_{threshold} & \text{用户属于描述动机} \\ M_u < M_{threshold} & \text{用户属于分类动机} \\ M_u = M_{threshold} & \text{用户动机待判} \end{cases} \tag{8}$$

临界值 $M_{threshold}$ 和判别系数 c 的确定方法参考文献[19]。$M_{threshold}$ 的计算如式(9)所示。

$$M_{threshold} = (n_1 \overline{M_d} + n_2 \overline{M_c})/(n_1 + n_2) \tag{9}$$

其中:n_1、M_d 为描述倾向用户数和所有描述用户 M 值的平均

值,n_2、M_c 为分类倾向用户数和所有分类用户 M 值的平均值。通过计算可以得到本文数据集的临界值 $M_{threshold}=0.4461$,所对应的判别系数为 $c=(0.854,0.713,0.496,0.261,0.091)$。

2.2　度量指标反映用户标注动机有效性检验

本文中用户动机倾向性的主要判定依据是 1.2 节中所提到的五种度量,那么这些指标是否真的有效,必须加以讨论。为了检验其有效性,笔者首先使用 OMDM 模型对样本用户进行动机倾向性判别,然后采取人工方式对样本用户的动机倾向性进行评判,进而比较两者的一致性。OMDM 模型判别使用式(8)即可,使用表 3 进行人工评判。样本数据来自于 Bibsonomy 和 Delicious 数据集。

表 3　不同动机的特征

	分类倾向	描述倾向
资源标签率	低	高
同义词出现情况	少	多
改变标签的代价	大	小
标签取自标题	少	多
词表规模	有限	无限
目的	浏览	查询与检索

人工评判过程如下:从 Delicious 和 Bibsonomy 中各选取三个用户,记为$\{P_1,P_2,P_3,P_4,P_5,P_6\}$。由这六个用户对他们每次标注按分类倾向和描述倾向进行归类,并完成从用户 A 的标签判断用户所属动机(表 4)。

表 4　从用户 A 的标签判断用户所属动机

	用户 A 的标签	分类倾向	描述倾向
URL_1	$t_1,t_2,\cdots,t_n$	□	□
URL_2	$t_1,t_2,\cdots,t_m$	□	□
⋮	⋮		
URL_n	$t_1,t_2,\cdots,t_u$	□	□
		□	□

本文采用平均 Cohen's Kappa 系数 κ 来验证人工评判与 OMDM 模型方法的一致性。如果 $0.61\leqslant\kappa<0.80$,可以认为人工评判与 OMDM 模型方法具有很好的一致性[20]。也就是说,五种指标能够真实反映用户动机。Cohen's Kappa 系数计算如表 5 所示,平均 Cohen's Kappa 系数 $\kappa=0.65$。

表 5　人工评判与 OMDM 模型方法的一致性检验

用户	P_2	P_3	P_4	P_5	P_6	OMDM
P_1	0.63	0.78	0.72	0.65	0.69	0.74
P_2		0.57	0.53	0.54	0.66	0.63
P_3			0.64	0.69	0.61	0.58
P_4				0.67	0.64	0.69
P_5					0.62	0.65
P_6						0.68

3　基于 OMDM 模型的标签推荐方法

3.1　基本思想及算法

基本思想是首先找到与待标注资源有相似动机倾向性的其他资源;然后筛选出与用户具有相似动机倾向性的资源,并聚合它们的标签;将这些标签作为候选推荐对象,依次计算候选推荐对象和待标注资源内容的相关性,将相关性大的标签推荐给用户。其算法描述如下:

输入:资源集合 R,用户集合 U,标签集合 T;特定用户 u;待标注资源$\hat{r}$。

输出:标签列表 L_u。

1　计算$\hat{r}$的五种度量指标,把 $\hat{r}$的动机倾向性表示为 $M_{\hat{r}}=(TRR_{\hat{r}},LFTU_{\hat{r}},TRCE_{\hat{r}},TSOF_{\hat{r}},STR_{\hat{r}})$。

2　计算 u 的五种度量指标,把 u 的动机倾向性表示为 $M_u=(TRR_u,LFTU_u,TRCE_u,TSOF_u,STR_u)$。

3　$R_{sim}=\Phi,\hat{T}_u=\Phi$

4　for i = 1 to $|R_u|$

5　计算 r_i 的五种度量指标,把 r_i 的动机倾向性表示为 $M_{r_i}=(TRR_{r_i},LFTU_{r_i},TRCE_{r_i},TSOF_{r_i},STR_{r_i})$。

6　for i = 1 to $|R_u|$

7　begin

8　$sim_{r,\hat{r}\in R_u}(M_r,M_{\hat{r}})=M_r\cdot M_{\hat{r}}/|M_r|M_{\hat{r}}$

9　if $sim_{r,\hat{r}\in R_u}(M_r,M_{\hat{r}})>\alpha$

10　$R_{sim}=R_{sim}\cup r$

11　end

12　for j = 1 to $|R_{sim}|$

13　begin

14　$sim_{r\in R_{sim}}(M_r,M_u)=M_r\cdot M_u/|M_r\|M_u|$

15　if $sim_{r\in R_{sim}}(M_r,M_u)>\beta$

16　$\hat{T}_u=\cup_{r|sim(M_r,M_u)\geqslant\beta}T_r$

17　end

18　$rec(t)=[0,\cdots,0],L_u=\Phi$

19　/* rec(t)为标签 t 与资源相关性 */

20　for each t in $\hat{T}_u$

21　for each w in $\hat{r}$

22　begin

23　$p(w)=\log(tf(w,r)/N_r+1)\log(N_R/|R(w)|+1$

24　$s(w,t)=\dfrac{\max(\{\log f(w),\log f(t)\})-\log f(w,t)}{\log N-\min(\{\log f(w),\log f(t)\})}$

25　计算 $rec(t)=\sum_w p(w)s(w,t)$

26　$L_u=L_u\cup t$

27　end

28　按照 rec(t)降序排列 L_u。

3.2　用户和资源的动机倾向性表示

正如引言所述,用户在标注时的动机倾向性并不是一成不变的,可能在标注某一资源时倾向于使用描述的标签,而换另一个资源时可能倾向于使用分类的标签。因此,用户的标注动机不可能很绝对地分成分类动机和描述动机,只能是倾向于某一种动机。基于 1.2 节提出的倾向性度量指标,本文把用户 u 的动机倾向性 M_u 表示为

$$M_u=(TRR_u,LFTU_u,TRCE_u,TSOF_u,STR_u) \tag{10}$$

其中:TRR_u、$LFTU_u$、$TRCE_u$、$TSOF_u$、STR_u 分别为用户 u 的五种度量指标。

在社会标签系统中,每一个资源被不同的用户在各种动机下进行标注。一个用户可能使用分类标签来标注,而另一个用户可能使用描述标签来标注同一资源。因此,可以认为不同类型的标签反映了资源的不同动机倾向性。那么,对于一个资源 r,同样可以用五种度量指标来表示其动机倾向性,如式(11)所示。

$$M_r=(TRR_r,LFTU_r,TRCE_r,TSOF_r,STR_r) \tag{11}$$

其中:TRR_r、$LFTU_r$、$TRCE_r$、$TSOF_r$、STR_r 分别为对资源的五种度量指标。

3.3 发现与待标注资源动机倾向性相似的资源

通常情况下以前使用过的标签是用户标注资源的首选。因此,可以充分利用与待标注资源相似的资源标签作为候选标签,这样将会得到与用户意图相似的标签。资源相似性计算采用基于向量空间的余弦法来计算,如式(12)所示。

$$\mathrm{sim}_{r,\hat{r}\in R_u}(M_r,M_{\hat{r}}) = M_r \cdot M_{\hat{r}} / |M_r| M_{\hat{r}} \tag{12}$$

其中:M_r、$M_{\hat{r}}$分别为用户已标注和待标注资源的动机倾向性表示。为了降低数据规模和找到更准确的标签集,本文仅考虑和待标注资源的动机倾向具有较高相似性的已标注资源。将这些已标注资源的集合表示为 R_{sim},即 $R_{\mathrm{sim}} = \{ r \mid \mathrm{sim}_{r,\hat{r}\in R_u}(M_r, M_{\hat{r}}) > \alpha \}$,$\alpha$ 为相似性控制因子。

3.4 计算 R_{sim} 中与用户动机倾向性相似的资源并聚合标签

计算 R_{sim} 中资源与用户倾向性的相关度,如式(13)所示。

$$\mathrm{sim}_{r\in R_{\mathrm{sim}}}(M_r,M_u) = M_r \cdot M_u / |M_r| \, \| M_u| \tag{13}$$

其中:M_r 为 R_{sim} 中资源的倾向性表示,M_u 为用户的倾向性表示。同样,为了降低数据规模和找到更符合用户意图的标签,这里只考虑与用户的动机倾向具有较高相似性的已标注资源。因此,设定控制因子 β 来控制聚合的资源数量,按照式(14)聚合与用户有着相似动机倾向性资源的全部标签,这个集合记为$\hat{T}_u$。

$$\hat{T}_u = \cup_{r \mid \mathrm{sim}(M_r,M_u) \geqslant \beta} T_r \tag{14}$$

3.5 生成推荐标签

资源的内容是标签推荐时不能忽略的重要因素。因此,当为用户标注特定的资源时推荐标签,必须考虑标签与特定资源内容的相关性。不同形式的资源刻画其内容的方式也不同。对于本文的网页资源,内容通过其文本来刻画。因此,对于网页资源,计算$\hat{T}_u$中标签与资源内容的相关性可以使用式(15)来计算。

$$P_{\mathrm{UMO}}(t \mid r) = \sum_{w\in R, t\in \hat{T}_u} p(w) s(w,t) \tag{15}$$

其中:$p(w)$是词 w 在资源 r 的内容中的权重,可以采用经典信息检索的 TFIDF 算法计算[21];而 $s(w,t)$则是词 w 和聚合标签集 T_u 中标签 t 之间的相关性,采用 Google 距离公式[22]来计算。

4 实验与分析

4.1 数据集

本文在两个数据集上进行了实验,Bibsonomy 数据集是来自于 2008 年 ECML/PKDD Discovery Challenge 竞赛,该数据集被认为具有描述倾向;第二个数据集是从 Delicious 抓取而来,该数据集被认为具有分类倾向。

4.2 实验设置

首先获取用户的动机倾向性,将每个数据集分为两部分,其中一部分用于获取用户的动机倾向性,另一部分用于评价推荐的质量,并采用文献[23]中基于 CRM 模型的推荐算法的推荐结果来考察基于 OMDM 模型的推荐算法;其次还要考察 α、β 对推荐准确性的影响。

4.3 评价方法

对于推荐系统来说,最客观的评价方法是在线实时统计用户从推荐候选标签中的选择比率。但由于客观现实不能满足,所以本文采用式(16)来计算推荐的准确率。

$$\text{准确率 } P = \frac{\mathrm{card}(\text{推荐标签列表} \cap \text{原始标签列表})}{\mathrm{card}(\text{推荐标签列表})} \tag{16}$$

表 6 给出了在 Bibsonomy 和 Delicious 数据集上基于 CRM 和 OMDM 两种模型进行推荐的准确率。从表 6 中可以看出,OMDM 的平均准确率超过了 CRM。分析这一现象的主要原因在于 OMDM 是从用户的标注动机出发进行推荐,因此推荐的标签更符合用户意图。

表 6 基于 CRM 和 OMDM 两种模型进行推荐的准确率比较

数据集	推荐方法	$P@5$	$P@10$	$P@20$
Bibsonomy	CRM	0.633	0.608	0.598
	OMDM *	0.756	0.713	0.705
Delicious	CRM	0.679	0.664	0.667
	OMDM *	0.792	0.787	0.792

表 7 为控制因子 α 对推荐准确性的影响。从表 7 可以看出,控制因子 α 对推荐准确性的影响比较大。分析其原因主要在于各个资源之间的倾向性差异较大,更深层的原因是资源的内容区别比较大。提升控制因子 α,推荐准确率增加明显。从表 7 中还可以看出,控制因子 α 对 Delicious 数据集的影响较 Bibsonomy 数据集大,这主要是因为 Bibsonomy 是对学术论文的标注系统,其多数标签是取自于学术论文的内容,而 Delicious 本身就是一个基于标签的分类系统。

表 7 控制因子 α 对推荐准确性的影响

数据集	控制因子	$P@5$	$P@10$	$P@20$
Bibsonomy	$\alpha=0.2$	0.639	0.627	0.594
	$\alpha=0.4$	0.644	0.648	0.609
	$\alpha=0.6$	0.661	0.671	0.615
	$\alpha=0.8$	0.704	0.711	0.693
	$\alpha=1.0$	0.713	0.720	0.704
Delicious	$\alpha=0.2$	0.625	0.588	0.591
	$\alpha=0.4$	0.736	0.646	0.653
	$\alpha=0.6$	0.812	0.793	0.778
	$\alpha=0.8$	0.837	0.824	0.817
	$\alpha=1.0$	0.903	0.886	0.843

表 8 为控制因子 β 对推荐性能的影响。从表 8 可以看出,控制因子 β 对推荐准确性的影响并不明显。分析其原因在于用户的动机倾向性和资源的动机倾向在一段时期内具有一定的稳定性。所以,提升控制因子 β,准确率只是略有增加。

表 8 控制因子 β 对推荐性能的影响

数据集	控制因子 β	$P@5$	$P@10$	$P@20$
Bibsonomy	$\beta=0.2$	0.707	0.679	0.629
	$\beta=0.4$	0.715	0.681	0.636
	$\beta=0.6$	0.726	0.718	0.682
	$\beta=0.8$	0.719	0.704	0.677
	$\beta=1.0$	0.734	0.729	0.697
Delicious	$\beta=0.2$	0.645	0.623	0.618
	$\beta=0.4$	0.656	0.631	0.625
	$\beta=0.6$	0.679	0.650	0.633
	$\beta=0.8$	0.696	0.663	0.652
	$\beta=1.0$	0.703	0.691	0.683

5 结束语

本文从用户标注动机不能绝对二分出发,提出用户在标注某个具体资源时实际上是存在不同的动机倾向,通过建立五种

度量指标,提出了基于动机倾向性的标签推荐模型(OMDM),并设计了基于该模型的算法。在两个不同的数据集上的实验结果表明,基于动机倾向性的社会标签推荐模型能获得更准确的备选标签。

本文在标签推荐研究上提供了一个新的研究视角,但研究本身还存在很多局限性。本文仅从五个观察指标对倾向性进行了度量,是否还存在其他不同质的度量指标,或者说从心理学、认知科学的理论角度有没有其他的指标来测量倾向性还有待研究。另外,本文提出的模型和算法并没有考虑所推荐的标签的新颖性、召回率,这也是笔者准备进行扩展的研究内容。

参考文献:

[1] SIGURBJÖRNSSON B, Van ZWOL R. Flickr tag recommendation based on collective knowledge[C]//Proc of the 17th International Conference on WWW. New York: ACM Press, 2008: 327-336.

[2] HOTHO A, JÄSCHKE R, SCHMITZ C, *et al*. Information retrieval in folksonomies: search and ranking[C]//Proc of the 3rd European Conference on the Semantic Web. Berlin: Springer, 2006: 411-426.

[3] SYMEONIDIS P, NANOPOULOS A, MANOLOPOULOS Y. A unified framework for providing recommendations in social tagging systems based on ternary semantic analysis[J]. IEEE Trans on Knowledge and Data Engineering, 2010, 22(2): 179-192.

[4] HARVEY M, BAILLIE M, RUTHVEN I, *et al*. Tripartite hidden topic models for personalised tag suggestion[C]//Proc of the 32nd European Conference on IR Research. Berlin Springer: 2010: 432-443.

[5] Motivation[EB/OL]. [2012-02-20] http://en.wikipedia.org/wiki/Motiva-tionJHJcite_note-0.

[6] TOM C. Two cultures of fauxonomies collide[EB/OL]. [2012-02-22]. http://www.plasticbag.org/archives/2005/06/two_cultures_of_fauxonomies_collide.

[7] HAMMOND T, HANNAY T, LUND B, *et al*. Social bookmarking tools (I): a general review [J]. D-Lib Magazine, 2005: 11(4).

[8] HECKNER M, HEILEMANN M, WOLFF C. Personal information management vs. resource sharing: towards a model of information behaviour in social tagging systems[C]//Proc of International AAAI Conference on Weblogs and Social Media. 2009: 42-49.

[9] WASH R, RADER E. Public bookmarks and private benefits: an analysis of incentives in social computing [J]. Journal of the American Society for Information Science and Technology, 2007, 44(1): 1-13.

[10] AMES M, NAAMAN M. Why we tag: motivations for annotation in mobile and online media[C]//Proc of SIGCHI Conference on Human Factors on Computing Systems. New York: ACM Press: 2007: 971-980.

[11] NOV O, NAAMAN M, CHEN Ye. Motivational, Structural and tenure factors that impact online community photo sharing [C]//Proc of the 3rd International AAAI Conference on Weblogs and Social Media. 2009: 138-145.

[12] SINHA R. A cognitive analysis of tagging [EB/OL]. [2012-02-24]. http://rashmisinha.com/2005/09/27/a-cognitive-analysis-of-tagging.

[13] MARLOW C, NAAMAN M, BOYD D, *et al*. HT06, tagging paper, taxonomy, Flickr, academic article, to read [C]//Proc of the 17th Conference on Hypertext and Hypermedia. New York: ACM Press, 2006: 31-40.

[14] XU Zhi-chen, FU Yun, MAO Jian-chang, *et al*. Towards the semantic Web: collaborative tag suggestions[C]//Proc of Collaborative Web Tagging Workshop at the WWW. New York: ACM Press, 2006.

[15] SEN S, LAM S K, RASHID A M, *et al*. Tagging, communities, vocabulary, evolution[C]//Proc of the 20th ANNIVERSARY Conference on Computer Supported Cooperative Work. Berlin: Springer, 2006: 181-190.

[16] STROHMAIER M, KÖRNER C, KERN R. Why do users tag? detecting users' motivation for tagging in social tagging systems[C]//Proc of ICWSM. 2010: 339-342.

[17] JIANG J J, CONRATH D W. Semantic similarity based on corpus statistics and lexical taxonomy [C]//Proc of International Conference on Research Computational Linguistics. 1997: 19-33.

[18] FU W T, KANNAMPALLIL T G, KANG Ruo-gu, *et al*. Semantic Imitation in social tagging[J]. ACM Trans on Computer-Human Interaction, 2010, 17(3): 1-37

[19] 范克新. 社会学定量方法[M]. 南京:南京大学出版社, 2004: 346-358.

[20] LANDIS J R, KOCH G G. The measurement of observer agreement for categorical data [J]. Biometrics, 1977, 33(1): 159-174.

[21] SALTON G, McGILL M J. Introduction to modern information retrieval [M]. [S. l.]: McGraw-Hill, 1983.

[22] CILIBRASI R, VITANYI P M B. The Google similarity distance [J]. IEEE Trans on Knowledge and Data Engineering, 2007, 19(3): 370-383.

[23] 靳延安, 李玉华, 刘行军. 不同粒度标签推荐算法的比较研究[J]. 计算机应用研究, 2012, 29(2): 504-509.

[24] 窦玉萌, 赵丹群. 协作标注系统研究综述 [J]. 现代图书情报技术, 2009, 3(2): 9-17.

(上接第 59 页)

[9] ZHOU Kang, GAO Zun-hai, XU Jin. An algorithm of DNA computing on 0-1 planning problem[J]. Advances in Systems Science and Applications, 2005, 5(4): 587-593.

[10] WANG Shi-ying, YANG Ai-ming. DNA solution of integer linear programming[J]. Applied Mathematics and Computation, 2005, 170(1): 626-632.

[11] ZHANG Feng-yue, YIN Zhi-xiang, LIU Bo, *et al*. DNA computation model to solve 0-1 programming problem[J]. Biosystems, 2004, 74(1-3): 9-14.

[12] 罗海波. 基于 0-1 规划的 DNA 计算模型的设计与实现[D]. 沈阳: 东北大学, 2008.

[13] 许进, 周康, 覃磊, 等. 0-1 规划问题的闭环 DNA 算法[J]. 系统工程与电子技术, 2009, 31(4): 947-951.

[14] BRAICH R S, CHELYAPOV N, JHONSON C, *et al*. Solution of a 20-variable 3-SAT problem on a DNA computer[J]. Science, 2002, 296(4): 499-502.

[15] SANCHES C A A, SOMA N Y. A polynomial-time DNA computing solution for the bin-packing problem[J]. Applied Mathematics and Computation, 2009, 215(6): 2055-2062.

[16] 孙伟, 尤加宁, 江宏, 等. 纳米粒子标记 DNA 探针的制备与检测应用[J]. 中国卫生检验杂志, 2005, 15(8): 1008-1010.