◎大数据与云计算◎

基于LDA和CTR的用户模型分析

吴飞飞,姬东鸿,吕超镇
WU Feifei, JI Donghong, LV Chaozhen

武汉大学 计算机学院,武汉 430072
Computer School of Wuhan University, Wuhan 430072, China

WU Feifei, JI Donghong, LV Chaozhen. Analysis of user model based on LDA and CTR. Computer Engineering and Applications, 2016, 52(6):50-54.

Abstract: Personal service is a hot topic. But how to construct an integrated user model remains a challenge for us. This paper makes use of the topic model LDA to infer the user model. In order to improve precision, CTR is put into use for restrict of feature vector. With a few manual factors, the machine generates a training topic model. Based on this model, 100 users' micro-log messages regarded as test data will be applied for evaluating the quality of recommendation. The results show that the recommendation of celebrity performs better than the recommendation of news. Generally speaking, personal service is satisfying.

Key words: Latent Dirichlet Allocation(LDA); topic model; Collaborative Topic Regression(CTR); user model; recommendation

摘 要: 个性化服务一直是研究的热点,但是如何构建完整的用户模型是一个颇有挑战性的问题。将基于主体模型LDA对用户模型进行预测,在用户和推荐项目的特征向量上采用CTR进行约束,使结果更为准确。在只需要少量人为因素下,由机器来训练最初的主题模型,在训练模型的基础上,通过选取100名用户的微博作为测试,用等级打分制来对推荐的项目进行打分,最终的结果显示,在新闻推荐上,微观满意度达到82.5%;而在名人推荐上,微观满意度达到了84.3%,综合以上,推荐服务的满意度还是令人满意的。

关键词: 隐形狄雷克雷分布(LDA);主题模型;基于主题模型的协同过滤(CTR);用户模型;推荐
文献标志码: A **中图分类号:** TP391 doi:10.3778/j.issn.1002-8331.1405-0179

1 引言

随着社交网络的流行,出现了大量的社交数据,微博是其中的代表。对于研究者来说,对微博进行分析是了解社交网络一个很好的渠道。在微博上,用户可以通过发布一段不超过指定长度(通常为140个字)的短文本来表达自己的观点[1],还可以分享自己的生活点滴,如果其他用户对某个用户感兴趣,可以关注他,并且转发或者回复他的微博来与他互动。正是因为如此,许多的研究者通过用户在社交网络上的互动记录来了解网络结构的变化[2-3]和信息的迁移[4-5],但是很少有人去研究怎样通过用户的这些网络活动来为用户提供个性化服务。提供个性化服务的前提是要构建一个好的用户模型,因此,本文就是通过分析个人微博来构建用户模型,一个完整的用户模型才会有一个好的个性化服务的体验。

2 研究现状

前人在这些问题上采用了一些方法,Zhao等[6]利用主题模型对推特与传统的在线媒体内容进行比较后发

基金项目: 国家自然科学基金重点项目(No.61133012);国家自然科学基金面上项目(No.61173062)。

作者简介: 吴飞飞(1988—),男,硕士生,研究领域为社交网络数据挖掘、个性化推荐等,E-mail:529828703@qq.com;姬东鸿(1968—),男,博士,博士生导师,研究领域为自然语言处理、语义网技术、机器学习、数据挖掘等;吕超镇(1990—),男,硕士生,研究领域为信息检索等。

收稿日期: 2014-05-15 **修回日期:** 2014-07-25 **文章编号:** 1002-8331(2016)06-0050-05
CNKI网络优先出版:2014-12-11, http://www.cnki.net/kcms/detail/11.2127.TP.20141211.1524.023.html

现，在推特上，人们喜欢谈论与家庭、生活相关的内容。Hong等[7]研究了在微博环境中怎样使用数据集来训练主题模型。Ramage等[8]利用Labeled-LDA对推特的内容和用户建模，在微博排序、用户推荐等方向上，都能够表现出不错的性能。Abel等人[9]通过提取微博中的实体、Hashtag等，把它们同当前的主流媒体如CNN、CBC、New York Times相链接，能够获取链接外的内容，从而拓展和丰富微博的语义，文献[10]中，提出了基于Twitter的用户模型应用TUMS。给定一个Twitter用户，收集该用户发布的所有微博，丰富语义，返回用户建模结果，并对其可视化。Michelson和Genc等将微博内容与维基百科资源相结合，并进行研究。

本文中，基于研究和评估，将对以下问题进行详细的描述：

(1)用户的兴趣随时间的迁移和受全局趋势的影响；

(2)用户模型的构建，考虑时间因素的影响；

(3)对个性化服务结果的评估。

为了回答以上三个问题，构建了一个框架来预处理微博消息。如图1所示，对微博消息进行停用词的处理，增加时间标签，对微博中包含的链接，抽取链接的正文内容[11-12]，也去除停用词。预处理完之后，将利用主体模型LDA对数据进行训练，在得到用户模型的过程中采用CTR对特征向量进行约束，最后用测试集对结果进行相应的评估，结果是令人满意的。

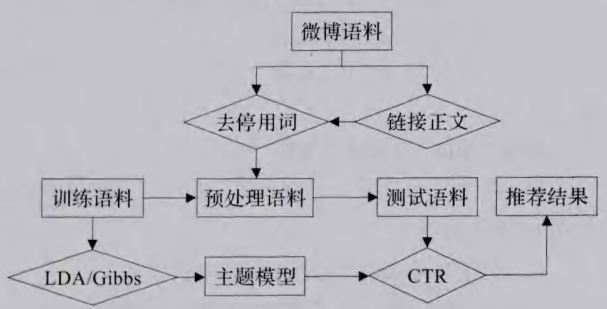

图1　框架流程图

3　用户兴趣的变化

为了更好地了解用户兴趣随时间的变化，分析了三名用户发布的微博主题情况，时间跨度是2013年10月1号到11月31号。把他们分别命名为user A，user B，user C，通过图2就可以看到他们的兴趣随时间的变化趋势。

通过图2，可以看出，一个人的兴趣是随着时间在变化的，在两个月的时间里，三名用户发布的微博主题都与其前面的主题不相同，甚至有些是不相关的。与此同时，还能够通过图中看出用户的兴趣也是受全局趋势影响的，例如在10月1号和11月11号的时候，微博数就都达到了他们这两个月微博数的峰值。一个是国庆假期，发布的几乎都是关于国庆的微博，另一个是光棍节，三名用户发布的几乎都是关于光棍和网购的微博。

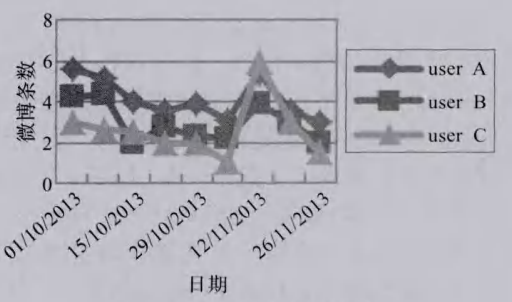

图2　用户微博主题变化图

4　用户兴趣模型

上章中讲到用户的兴趣会随着时间而变化，因此不同的词语有不同的权重，即使是相同的词语在不同时间也有不同的权重。因此构建用户模型的时候，采取了用户-时间-主题三维模型[13]，该模型是在传统的用户-物品的二维基础上加入了时间因素这一维。传统的用户-主题二维模型，最主要的是构建用户模型，对于某个用户来说，用户的微博消息会涵盖各种主题，在这种背景下，以用户所发微博的特征来描述用户的特征，对每个词汇也会加上权衡因子。对词汇不仅考虑了它的频率问题，与此同时加入了时间因素。给某个词汇赋予相应的时间权重，时间与当前离得越远，权重就越小；反之则越大。

定义1(淡化因子) 由于用户的兴趣随时间逐渐消退和变化，相应的影响因子就会衰减。

$$w(c,time,T_{\text{weibo},u})=\sum_{t\in T_{\text{weibo},u,c}}\left(1-\frac{|time-time(t)|}{max_{\text{time}}-min_{\text{time}}}\right)^{d} \quad (1)$$

表达式中，$T_{\text{weibo},u,c}$表示用户u发布的微博的集合和参考的概念。$Time(t)$表示返回给定微博t的时间戳，max_{time}和min_{time}分别表示发布微博的最远时间和最近时间，参数d用来调节时间距离的影响。d设置得越大，距离输入时间越远的概念的惩罚越重，相应的分值就比时间距离短的低很多。

4.1　主题模型LDA

本文中选取的主题模型是LDA，采用吉布斯采样估计LDA参数，用CTR对特征向量进行约束。首先在这里简单介绍一下LDA。图3就是LDA模型。α和β表示语料级别的参数，选定一个主题向量θ，确定每个主题被选择的概率。然后在生成每个单词的时候，从主题分布向量θ中选择一个主题z，按主题z的单词概率分布生成一个单词。z由θ生成，w由z和β共同生成。

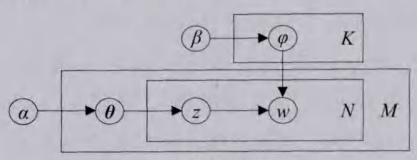

图3　主题模型LDA

LDA是一种非监督的学习模型，一般被用来作为识别大规模文档集或者语料库中潜在主题信息的工具。它采用了词袋的方法，这种方法将每一篇文档视为

一个词频向量,从而将复杂难懂的文本信息转化为了易于建模和可量化的数字信息。但是这种方法并没有考虑词与词之间的顺序,这也就是说把问题的复杂性降低了,同时也为模型的改进提供了契机。每一篇文档代表了一些主题所构成的一个概率分布,而每一个主题又代表了很多单词所构成的一个概率分布。对于每一个主题 Z,把出现在 Z 中的词语 W 的频率与主题 Z 下的文档 D 中其他的词语数目相乘就得到该词语的一个概率。

$$P(Z|W,D)=\frac{k+\beta_w}{N+\beta}\times(n+\alpha) \tag{2}$$

k 表示主题 Z 中 W 词的出现次数,N 为主题 Z 中所有的词数目,n 为主题 Z 下文档 D 的词语数目,α,β,β_w 被称为超参数,作为平滑因子存在。

4.2 吉布斯采样

在最初提出LDA模型的时候采用的是EM来计算,现在普遍采用更为简单但有效的吉布斯采样,迭代完成后输出主题-词参数矩阵 $\boldsymbol{\Phi}$ 和文档-主题矩阵 $\boldsymbol{\theta}$。

$$\Phi_{k,t}=\frac{n_k^{(t)}+\beta_t}{\sum_{t=1}^{V}n_k^{(t)}+\beta_t} \tag{3}$$

$$\theta_{m,k}=\frac{n_m^{(k)}+\alpha_k}{\sum_{k=1}^{K}n_m^{(k)}+\alpha_k} \tag{4}$$

m 表示第 m 篇文档,n 表示文档中的第 n 个词,k 表示主题,K 表示主题的总数,$(n_k^{(t)}+1)$ 表示 k 主题对应的 t 词的次数,$(n_m^{(k)}+1)$ 在 m 文档中 k 主题出现的次数。

4.3 用户兴趣识别

因为微博较短,而且语法也不太规范,对于特征的描述也比较困难。在这里,对于用户兴趣主要是从词频和主题上进行描述。词频用微博统计和卡方分布相结合。主题就是基于大规模微博数据训练主题模型,对微博数据的主题进行推断,以各微博在主题上的分布作为特征,设置主题类别为 K 个,对应的主题为 $C=\{C1,C2,\cdots,Ck\}$,用户 U 在一段时间内发布的微博数目为 n 条,根据事先训练的主题模型进行预测,就可以得到每一条对应的主题概率为 $Pi=\{Pi1,Pi2,\cdots,Pik\}$,在对每一条微博进行预测的同时,加入了时间因素,也就是前面说的淡化因子,时间越早的微博对用户模型的影响就越小,反之则越大。根据LDA主体模型的预测,每一条微博都会产生 K 个概率,显示出来该条微博所属的主题情况。不能简单地把每条微博相应主题的概率加起来,最后再比较它们的大小,选择最大的作为该用户最感兴趣的主题。

$$p(C_i)=\sum_{j=1}^{n}p_{ij}\times w(c,time,T_{\text{weibo},u}) \tag{5}$$

4.4 协同过滤算法CTR

CTR(Collaborative Topic Regression)是一种基于主题模型的协同过滤算法,它集合了传统的协同过滤算法[14-15]和主题模型的优点。比如对于文章A和文章B都是讲关于社交网络中的机器学习的应用,在主题分布中有 θ_A 和 θ_B,进一步假设这两篇文章对应不同群体,A描述的是在社交网络应用中的机器学习,而B应用了标准的机器学习,但重点在社交网络数据的分析上。但是在主题分布上A和B极其相似,都会推荐给用户,所以在这里运用协同过滤,CTR能发现两者的不同之处,而改善推荐效果。因此,CTR用来约束用户和推荐项目,也就是说对用户的特征向量和推荐项目的特征向量进行约束,使它们都符合以某种主题分布向量为均值的正态分布。图4是CTR模型的概率图。

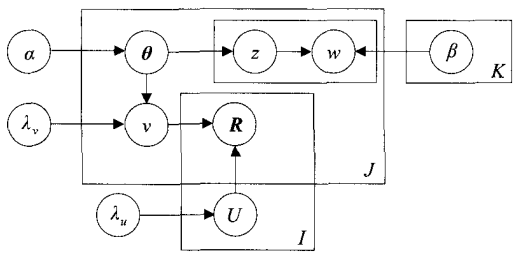

图4 CTR框架图

CTR模型既有传统的协同过滤算法的简单性,又有概率主题模型的优点。用于表示用户的特征向量,它是符合均值为0的正态分布,用于表示用户的兴趣;而推荐项目的特征向量符合均值为 $\boldsymbol{\theta}$ 的正态分布,其潜在方差为 ε。下面就是CTR模型的运行过程。

CTR模型的运行过程:

输入 用户的正则化系数 λ_u,推荐项目的正则化系数 λ_v。

输出 矩阵 $\boldsymbol{R}$ 的逼近矩阵 $\boldsymbol{X}$。

(1)对于每个用户 i,从分布 $N(0,I_k/\lambda_u)$ 中抽取用户的潜在特征向量 $\boldsymbol{U}_i$,即 $\boldsymbol{U}_i\sim N(0,I_k/\lambda_u)$。

(2)对于每个文本形式的推荐项目 j:

①运用前面的LDA模型得到主题分布 θ_j。

②得到推荐项目的潜在方差 ε_j,ε_j 满足分布 $N(0,I_k/\lambda_v)$。

③得到推荐项目的特征向量 $\boldsymbol{V}_j=\theta_j+\varepsilon_j$。即 $V_j\sim N(\theta_j,I_k/\lambda_v)$。

(3)对于每个评分点 (i,j),得到相应的预测评分 $x_{ij}\sim N(\boldsymbol{U}_i^{\mathrm{T}},C_{ij}^{-1})$。

在这里参数 C_{ij} 是对于评分点的值 X_{ij} 的信任参数,参数 C_{ij} 的值越大,表示评分值 X_{ij} 越可信;当 C_{ij} 为0时,就说明用户 i 对这个推荐项目没有兴趣或者根本就没有注意到项目 j。与文献[12-13]一样,为参数 C_{ij} 赋予一定的权值。

$$C_{ij}=\begin{cases}a, X_{ij}=1\\ b, X_{ij}=0\end{cases}$$

这里的 a,b 是控制参数,满足 $0<b<a\leqslant 1$。

在CTR模型中，对推荐项目 j 的特征向量进行了一定的约束，使其满足均值为推荐项目 j 的主题分布向量的正态分布，即 $V_j \sim N(\theta_j, I_k/\lambda_v)$；同理，也对用户的特征向量实施了约束，它也符合某种正态分布，即 $U_i \sim N(\theta_i, I_k/\lambda_u)$。

5 实验与分析

5.1 实验数据

数据集是由新浪微博和武汉大学自然语言处理实验室合作的项目提供的，包含多达用户50 000人，微博条数超过2 700万。训练数据包含用户50 000人，微博27 398 122条；测试数据包含用户100人，微博条数达到52 956条，具体的数据分布如表1。

表1　训练数据分布表

项目	年龄				
	15~25	26~35	36~45	46~56	>56
人数	19 603	14 729	10 038	5 472	158
链接数	1 068 913	756 358	463 255	258 648	7 691
微博数	12 233 170	8 191 203	4 908 350	2 029 017	36 382
微博均值	5.19	4.63	4.07	3.09	1.92

表1是训练数据，训练数据中的用户是基于新浪微博提供的数据集，统计了他们的年龄分布，微博数以及微博密度。首先对数据进行了简单的处理，去掉了停用词，对于包含外部链接的微博，把外部链接的正文部分抽取出来也去掉了停用词。

表2是测试数据，对于测试数据中的用户，为了得到与训练数据同样的年龄分布，发动了身边的同学，朋友以及老师，共100位。对于这些用户收集和分析他们的微博，推测出他们的兴趣所在。在分析了他们的微博，得出了用户模型，经过算法推荐给他们相关的新闻和名人，采用问卷形式，对于每个项目都推荐20条，最终根据用户自己的主观喜好来判断推荐的事物是否是他们所喜欢的，根据不同的喜好程度对各项推荐进行打分，这也是采集分数的方法。对测试数据进行预处理，然后根据训练出来的主题模型，根据CTR的约束得到它的预测主题分布，计算出它的推荐项目。

表2　测试数据分布表

项目	年龄				
	15~25	26~35	36~45	46~56	>56
人数	37	27	22	11	3
链接数	2 445	1 102	1 003	523	177
微博数	23 091	14 905	10 573	3 679	708
微博均值	5.20	4.60	4.00	2.78	1.96

5.2 评价策略

采用的评价策略是让具体的用户对推荐的结果进行打分，根据最后的分数，就能知道推荐效果的满意度。具体的公式如下：

$$\text{宏观满意度} = P/N \tag{6}$$

$$\text{微观满意度} = K/N \tag{7}$$

P 表示总得分，N 表示总项目数；K 表示标记为喜欢或以上的项目数。除了上面的评价标准，还引入了传统的评价体系——准确率 P、召回率 R、F 值，具体表达式如下：

$$P = X/M \tag{8}$$

$$R = Y/M \tag{9}$$

$$F = (2 \times P \times R)/(P + R) \tag{10}$$

X 表示用户满意的推荐项目数，Y 表示正确属于该类的推荐项目数，M 表示向用户推荐的项目总数。

采取了分级制，如表3，给不同的满意度不同的分数，分别对新闻和名人的喜欢程度进行打分。

表3　分级制相应的分数和态度

项目	分数			
	1	2	3	4
新闻	不喜欢	可能喜欢	喜欢	非常喜欢
名人	不喜欢	可能喜欢	喜欢	非常喜欢

在两项推荐项目中，对于新闻推荐，有100个用户，分别对每个用户推荐20篇新闻；而对于名人推荐，向每位用户推荐20个名人，根据他们的个人情况，对推荐的新闻和名人进行打分，最终的结果显示如表4。

表4　得分情况统计表

项目	新闻	名人	各等级得分数
得1分数目	144	218	362
得2分数目	206	95	602
得3分数目	771	304	3 225
得4分数目	879	1 383	9 048
总分	6 385	6 852	13 237
平均分	3.192 5	3.426 0	3.309 2
宏观满意度	0.798 1	0.856 5	0.827 3
微观满意度	0.825 0	0.843 5	0.834 2

从表4中可以看出，在名人推荐上的满意度要比新闻推荐好，这是因为人们发的有关名人微博时，一般都说得比较清楚，对于喜欢的明星，人们才会有兴趣发布有关他们的信息，当然也不排除有些发布的是不喜欢的名人的微博，所以名人推荐中，不喜欢这一类也占有10.9%。虽然在宏观满意度上只有接近80%的满意度，但更看重的是微观满意度，因为微观满意度更能体现推荐的效果。

与此同时，有另外4组实验来对实验结果进行补充，分别是：(1)使用了基于概率的潜在语义模型，表示为PLSA；(2)使用了LDA没有使用CTR也没有使用时间因素的，表示为LDA；(3)使用LDA和时间因素但是没有使用CTR的，表示为(LDA+time)；(4)使用LDA并且使用CTR但没有使用时间因素，表示为(LDA+CTR)。分别从宏观满意度、微观满意度、准确率、召回率以及 F 值进行比较，实验结果如表5。

表5中显示，因为准确率和微观满意度的定义重复，因此它们的值相等。在这五种算法中，第五种，也就

表5 实验结果对比表

评价算法	宏观满意度		微观满意度		P		R		F	
	新闻	名人	新闻	名人	新闻	名人	新闻	名人	新闻	名人
PLSA	0.552 4	0.623 9	0.559 2	0.630 2	0.559 2	0.630 2	0.588 4	0.688 6	0.573 4	0.658 1
LDA	0.592 5	0.653 3	0.601 1	0.649 8	0.601 1	0.649 8	0.691 4	0.723 5	0.643 1	0.684 7
LDA+time	0.655 3	0.672 5	0.659 7	0.670 2	0.659 7	0.670 2	0.702 5	0.744 7	0.680 4	0.705 5
LDA+CTR	0.679 1	0.711 8	0.701 3	0.722 0	0.701 3	0.722 0	0.706 2	0.752 3	0.703 7	0.736 8
LDA+CTR+time	0.798 1	0.856 5	0.825 0	0.843 5	0.825 0	0.843 5	0.921 3	0.978 2	0.870 5	0.905 9

是本文提出来的算法,它的整体表现是最好的。所有的算法在名人项目上表现都比在新闻上要好,这是因为对于关注的名人,结果很清晰,对于新闻种类繁多,并不能从模型中体现出各个方面。结合了时间因素和添加了CTR的LDA模型,两者的结果表现也比单纯的LDA要好。单独的LDA与PLSA的表现差不太多。

在不同个数的主题下,表现也各不一样,通过分析发现在170~220个主题下,推荐的效果是最好的,因为当主题数太少,分布就太笼统,推荐的物品也会比较抽象,几乎大部分的东西都是差不多的主题分布;而当主题数太大,就表示物品被分得很细,但是当某个人的用户模型被太多的标签描述的时候,主题就会太分散,也不利于用户模型的构建,所以结果如图5所示。

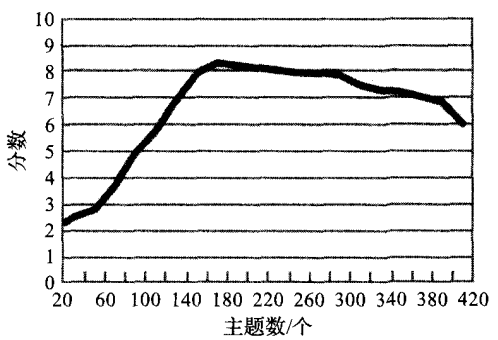

图5 不同个数主题下的推荐满意度

6 结束语

本文中,用到了时下比较流行的主体模型LDA,采用了吉布斯采样,并用CTR对特征向量进行约束,对主题的预测进行了多方面的评价,在LDA的基础之上,考虑了时间因素对用户模型的影响,而且结果显示,时间因素的影响还是蛮大的。在评价阶段,不同方面的推荐效果还是有一定的区别,说明要构建一个完整的用户模型,能适用到各个方面是有一定困难的,但这也是往前钻研的动力,遇到困难,解决困难,科学才有进步,提出更好的主题模型或者有更好的算法,个性化服务的满意度才会提高。

参考文献:

[1] 宋巍,张宇,谢毓彬,等.基于微博分类的用户兴趣识别[J].智能计算机与应用,2013,3(4):80-83.

[2] Cha M, Haddadi H, Benevenuto F, et al.Measuring user influence in Twitter: the million follower fallacy[C]//ICWSM, 2010:10-17.

[3] Weng J, Lim E P, Jiang J, et al.Twitter rank: finding topic-sensitive influential winterers[C]//Davison B D, Suel T, Craswell N, eds.Proceedings of the Third International Conference on Web Search and Web Data Mining(WSDM), New York, NY, USA, 2010:261-270.

[4] Kwak H, Lee C, Park H, et al.What is twitter, a social network or a news media?[C]//Proceedings of the 19th International Conference on World Wide Web(WWW), 2010:591-600.

[5] Lerman K, Ghosh R.Information contagion: an empirical study of the spread of news on digg and twitter social networks[C]//Proceedings of 4th International Conference on Weblogs and Social Media(MedialCWSM), 2010.

[6] Zhao W X, Jiang J, Weng J, et al.Comparing twitter and traditional media using topic models[M]//Advances in Information Retrieval.Berlin Heidelberg: Springer, 2011: 338-349.

[7] Hong L, Davison B D.Empirical study of topic modeling in Twitter[C]//Proceedings of the SIGKDD Workshop on SMA, 2010.

[8] Ramage D, Dumais S, Liebling D.Characterizing microblogs with topic models[C]//International Conference on Weblogs and Social Media, 2010.

[9] Abel F, Gao Qi, Jang.Semantic enrichment of Twitter posts for user profile construction on the social Web[C]//Weblogs and Social Media, ICWSM, 2010.

[10] Abel F, Gao Qi, Jang.TUMS: Twitter-based user modeling service[C]//ESWC, 2011.

[11] 孙承杰,关毅.基于统计的网页正文信息抽取方法的研究[J].中文信息学报,2004,18(5):17-22.

[12] 熊忠阳,蔺显强,张玉芳,等.结合网页结构与文本特征的正文提取方法[J].计算机工程,2013,39(12):200-203.

[13] Campos P G, Díez F, Cantador I.Time-aware recommender systems: a comprehensive survey and analysis of existing evaluation protocols[J].User Modeling and User-Adapted Interaction, 2014, 24(1).

[14] Wang C, Blei D.Collaborative topic modeling for recommending scientific articles[C]//ACM KDD, 2011:448-456.

[15] Purushotham S, Liu Y, Kuo C C J.Collaborative topic regression with social matix factorization for recommendation systems[C]//ACM ICML, 2012:1255-1265.