考虑用户–发布者关系的个性化微博搜索模型

张永棠[1,2,3]，罗海波[1,2]

(1. 广东东软学院计算机科学与技术系 广东 佛山 528225；2. 广东省大数据分析与处理重点实验室 广州 510006；
3. 南昌工程学院江西省协同感知与先进计算技术研究所 南昌 330003)

【摘要】提出了一种考虑用户与发布者建模的个性化微博搜索模型，该模型一方面运用主题模型与语言模型构建微博主题维度的用户兴趣模型，另一方面，融合用户与微博发布者的关系特征，构建用户–发布者关系维度的用户兴趣模型。并将二者进行有效融合，设计了将单个用户的微博作为一个文本的训练方法，解决微博文本短、语料稀疏的问题。基于真实用户搜索反馈的实验表明，融合用户–发布者关系的微博搜索模型可有效提高微博搜索的个性化效果。

关 键 词 信息搜索； 发布者模型； 关系模型； 社交网络； 主题模型
中图分类号 TP301 **文献标志码** A doi:10.3969/j.issn.1001-0548.2018.04.024

Personalized Micro-Blog Search Model Considering User-Publisher Relationship

ZHANG Yong-tang[1,2,3] and LUO Hai-bo[1,2]

(1. Department of Computer Science and Technology, Guangdong Neusoft Institute Foshan Guangdong 528225;
2. Guangdong Key Laboratory of Big Data Analysis and Processing Guangzhou 510006;
3. Institute of Cooperative Sensing and Advanced Computing Technology, Nanchang Institute of Technology Nanchang 330003)

Abstract A personalized micro-blog search model is proposed by taking into account the user and publisher's modelling. In this search model, the user interest model of the micro-blog theme dimension is constructed by using the theme model and language model, while the user interest model of the user-publisher relationship dimensions is constructed by integrating the features of user-publisher relationship. These user interest models are further integrated to design a single user's micro-blog as a text of the training methods for solving the problems of short texts and sparse notes of the micro-blog. Experiments based on real users' search feedback show that the proposed micro-blog search model can improve the personalized effect of micro-blog search.

Key words information search; publisher model; relational model; social networks; topic model

微博已经成为人们信息沟通的重要渠道。截至2016年底，中国微博用户数量已达到3.906亿，每天的微博发布总量超过一亿[1]。微博给用户提供了更丰富的信息的同时，也带来了大量冗余信息，因此，用户通过微博平台获取精准的信息越来越难。

目前面向微博的信息搜索研究主要聚焦于如何将微博区别于网页的特征引入微博搜索排序模型，从而改进微博搜索效果。如，文献[2]提出了基于微博特征的微博内容重要性计算模型，并利用协同过滤方法进行微博内容的个性化推荐。文献[3-4]构建了考虑时间因素的微博内容排序模型。文献[5]设计了一种用户查询和微博内容的关系度量方法，提出了一种面向微博的查询拓展模型。文献[6]将社交网络信息引入微博搜索建模，利用用户朋友的搜索兴趣构建微博用户的兴趣建模。针对微博内容短、主题广的特点，文献[7]利用主题模型构建用户兴趣模型，文献[8]提出了基于文献[9]的微博搜索技术。

上述研究虽然将微博特征引入微博搜索模型，对面向网页的搜索进行了扩展。但是，在进行用户兴趣建模时，并未考虑用户–发布者关系这一关键要素。因此，本文提出了一种考虑微博内容、发布者特征和用户–发布者关系的微博个性化搜索模型。

1 准备知识

1.1 语言模型和主题模型

语言模型在信息检索领域表现出色，它的简洁

收稿日期：2017 – 01 – 09；修回日期：2017 – 08 – 24
基金项目：国家自然科学基金(61363047)；广东省大数据分析与处理重点实验室开放基金(2017007)
作者简介：张永棠(1981 –)，男，副教授，主要从事光通信系统、网络空间安全方面的研究.

灵活也适用于处理一些复杂的新型检索问题。查询似然模型是语言模型中的经典方法，其思路是为每个文档构建一个语言模型 $\boldsymbol{M}_d$，根据文档模型生成一个查询的概率 $P(\boldsymbol{D})$ 对文档进行排序[10]，有：

$$P(\boldsymbol{D}\,|\,\boldsymbol{Q})=P(\boldsymbol{Q}\,|\,\boldsymbol{D})P(\boldsymbol{D})/P(\boldsymbol{Q}) \quad (1)$$

式中，$P(\boldsymbol{D})$ 为先验通常融合权威度、新颖度、文本长度等信息。将文档与查询视为一元模型，则一个文本模型生成查询的概率为：

$$\hat{P}_{\mathrm{LM}}(q\,|\,\boldsymbol{M}_d)=\prod_{t\in q}\hat{P}_{\mathrm{mle}}(t\,|\,\boldsymbol{M}_d) \quad (2)$$

式中，$\hat{P}_{\mathrm{mle}}(t\,|\,\boldsymbol{M}_d)$ 为语料集生成词 t 的极大似然估计。

语言模型中的重要问题是如何处理未出现的词，Jelinek-Mercer(JM)平滑[11]是常用的方法。JM平滑用整个语料中一个词出现的概率来代替该文档中未出现词的概率：

$$P(q\,|\,\boldsymbol{M}_d)=\lambda P_{\mathrm{LM}}(q\,|\,\boldsymbol{M}_d)+(1-\lambda)P_{\mathrm{LM}}(q\,|\,\boldsymbol{M}_{\mathrm{corpus}}) \quad (3)$$

语言模型作为一种无监督的文本分析方法，可将文本根据其隐含的主题进行分类。Latent Dirichlet Allocation(LDA)是一种流行的主题模型，Author-Topic模型作为LDA的一种扩充将作者引入了主题模型之中。由于主题模型粒度过大，LDA模型被作为平滑策略引入LM[11]有：

$$P(q\,|\,\boldsymbol{M}_d)=\lambda P_{\mathrm{TM}}(q\,|\,\boldsymbol{M}_d)+(1-\lambda)P_{\mathrm{LM}}(q\,|\,\boldsymbol{M}_d) \quad (4)$$

微博具有文本短、主题丰富的特性，运用Open Directory Project无法很好地覆盖用户感兴趣的主题，而主题模型可以解决该问题，将用户感兴趣的主题映射在主题模型上。由于微博的短文本特征，关键词可能只出现一次，会对真正有用的词造成较大影响；若只使用主题模型，由于主题词的粒度大，无法很好展现用户在主题上的兴趣。主题模型与语言模型的混合可有效表征用户在主题上的兴趣，通过二者的融合，将词赋予其主题，并和主题内其他词关联，可有效解决查询歧义问题。

1.2　质量模型与作者模型

针对语言模型的研究中，人们主要关注文档模型生成查询的概率，文档的先验概率常被视为相同而忽略，至今没有统一的方法来为文档质量、文档作者这些可表征文档先验概率的特征建模。文献[12]提出4个要素来建立微博读者的信任模型：博主的经验与身份、博主的可信度和价值观、信息质量、博主的魅力和个性。文献[13]通过线性加权方式融合两个方面的11种信息来为博客构建信任模型：博文信息(大写、表情、强调性内容、错误拼写、文本长度、时间因素、语义)，博文可信度(垃圾过滤、评论、正规度、主题一致性)。文献[14]首次在微博搜索中融入排序策略和社交因素，从发布者的角度分析发布的微博数及发布者的入度出度情况；并从微博质量角度考虑文本长度与URL。文献[15]利用SVM算法，使用微博的良构性[16]、事实性、导航质量训练构造器来对微博质量进行评估。文献[17]从URL、用户名和标签、感叹词、正负向词语、表情、情感、词语、主题来为微博建立模型。本文则从主题的角度进行思考，挑选对于用户选择微博较为关键的要素，从用户与微博作者的兴趣关联、该条微博是否属于作者所专注的主题、该微博本身的质量3个角度来进行个性化建模，并和语言模型与主题模型的混合模型相融合。

2　考虑作者建模的微博个性化搜索

本文通过微博内容建模、用户兴趣建模、发布者特征建模和用户-发布者关系建模等步骤，构建面向微博的个性化搜索模型。

2.1　用户兴趣建模

表1　用户微博中词所属的主题

用户发表的微博	主题
林俊杰[1]新曲[1]MV[1]邀请滨崎步[1]，期待！ 小米[2]升级[2]了V5[2]之后，依然流畅[2] 周杰伦[4]要结婚[4]了，一直听[1]周杰伦[1]的歌[1] 据说魅族[2]用户[2]忠诚度[2]仅次于小米[2] 烟台[3]苹果[3]真是好吃[3]	1：音乐 2：科技数码 3：美食 4：娱乐八卦

表2　用户在各个主题下的概率与词频

$P(T1)$=0.025 $T1$：音乐		$P(T2)$=0.420 $T2$：科技数码		$P(T3)$=0.036 $T3$：美食		…	
w	$c(w)$	w	$c(w)$	w	$c(w)$	w	$c(w)$
林俊杰	4	小米	12	热干面	8	…	…
MV	2	魅族	8	武汉	2	…	…
滨崎步	2	升级	5	苹果	1	…	…

本文融合语言模型与主题模型对用户兴趣进行建模。如表1、表2所示，主题层展现了用户在主题粒度上的偏好，而语言模型则展现了用户更喜欢某一主题下的哪一个词，语言模型与主题模型的双层模型结构可以更好地反映用户兴趣。

为了构建用户兴趣的双层模型，在主题层将用户u所发布的所有微博视为一个文档。由于用户感兴趣的主题往往多种多样，本文提出一种改进的Author-Topic模型进行文本主题训练。与传统Author-Topic模型[15]不同，本文采集了大量用户的微博作为语料集进行LDA训练，得到了通用的主题

模型。基于训练好的主题模型，使用Gibbs抽样[12]来获取用户兴趣在主题下的分布：

$$p(z_w = i) \propto p(\phi_i \mid \boldsymbol{D}) p(z_w \mid \phi_i) \tag{5}$$

式中，w为文档$\boldsymbol{D}$中的一个词，z_w为该词的主题；右侧的概率分别对应着Dirichlet后验分布在贝叶斯框架下的参数估计，而此时的$p(z_w)$来自训练好的模型。得到每一个词的分布之后，统计该文档的Topic-Word共现概率矩阵，即可推导出用户所有微博在各主题上的分布$\theta_{u,k}^{\mathrm{IM}}$。

在词级别，由于不能保证一条微博中只包含一种主题信息，所以本文利用训练好的主题模型得到用户的主题-词矩阵，使用每一个主题下的词来构建语言模型：

$$\theta_{u,k,w}^{\mathrm{IM}} = \frac{\sum c(w)}{\sum_{w' \in \boldsymbol{V}} \sum c(w')} \tag{6}$$

式中，IM为用户模型；$c(w)$为该主题下w词的数量；$\boldsymbol{V}$为词典。

针对未出现的新词，本文将训练好的模型的主题-词分布融合进语言模型，对其进行平滑：

$$\hat{\theta}_{u,k,w}^{\mathrm{IM}} = (1-\lambda)\theta_{u,k,w}^{\mathrm{IM}} + \lambda P(\omega \mid f_k^{\mathrm{TM}}) \tag{7}$$

式中，λ为Jelinek-Mercer平滑参数。进一步融入用户在主题层的兴趣，可得到：

$$\hat{\theta}_{u,k,w}^{\mathrm{IM}} = (1-\lambda)\theta_{u,k,w}^{\mathrm{IM}} \theta_{u,k}^{\mathrm{IM}} + \lambda P(\omega \mid f_k^{\mathrm{TM}}) \tag{8}$$

虽然发表的微博是用户兴趣的直接体现，但用户发布的微博数量往往较少，基于用户发布的微博构建的用户模型往往无法完全反应用户兴趣。从文献[15]可以知道，结合用户社交关系信息可以提高用户兴趣模型的有效性。

本文将社交关系融入用户兴趣模型，假设用户关注的朋友们为$\boldsymbol{F}$，利用$\boldsymbol{F}$中用户发布的微博内容对IM进行修正。实际应用中，在一个用户的所有朋友中，与用户具有相似兴趣的往往只占很少比例。而且，用户在关注某一朋友时，往往是对他发表的某一方面的微博感兴趣。本文利用交互次数、兴趣相似度和影响力对用户朋友进行排序，将朋友发布的相似微博内容提取出来，融入用户兴趣模型，引入以下指标[6]。

1) 交互次数：用户转发、提及、交互的次数越多，说明用户之间的兴趣交叉点越多。首先将用户的转发、提及的微博抽取出来，进行Gibbs抽样得到主题-词矩阵，再对每一个博主的各个主题的词进行汇总，得到用户对博主在主题上的兴趣为：

$$\omega_I(u,f,k) = \frac{c(f,k)}{c(I,k)} \tag{9}$$

式中，$c(I,k)$为该用户转发评论中属于主题$\boldsymbol{K}$的词；$c(f,k)$为转发该博主微博中属于主题$\boldsymbol{K}$的词。

2) 兴趣相似度：使用KL-divergence[6]来衡量用户与朋友之间的相似度：

$$\omega_S(u,f) = 1/\mathrm{KL}(\theta_u \| \theta_f) \tag{10}$$

3) 影响力：使用朋友的粉丝数来评价用户的影响力，并将其标准化：

$$\omega_p(f) = \lg(\mathrm{popularity}) / \lg(\max) \tag{11}$$

式中，max为微博中最大的粉丝数。

此外，由于用户对主题的偏好存在差别，将主题偏好$\omega_T(u,k)$作为先验融入到模型中。此时，$\boldsymbol{F}$的权重可定义为：

$$\omega_{u,f,k} = \boldsymbol{\sigma}^{\mathrm{T}} \begin{pmatrix} \omega_p(f) \\ \omega_I(u,f,k) \\ \omega_S(u,f) \\ \omega_T(u,k) \end{pmatrix} \tag{12}$$

式中，$0 \leqslant \omega_{u,f,k} \leqslant 1$；$\boldsymbol{\sigma}$是权重向量用于调节不同部分的影响。之后本文将所有朋友的权重标准化$\sum_{f \in \boldsymbol{F}_u} \omega_{u,f,k} = 1$，$\boldsymbol{F}_u$是用户的所有朋友，只挑选用户朋友中评分最高的10人进行计算。

此时，基于文献[6]的协同模型，可得到用户的朋友模型CM：

$$\theta_{u,k,w}^{\mathrm{CM}} = \sum_{f \in \boldsymbol{F}_u} \omega_{u,f,k} \theta_{f,k,w}^{\mathrm{IM}} \tag{13}$$

将朋友模型CM融合进用户模型IM，可得：

$$\hat{\theta}_{u,k,w}^{\mathrm{IM,CM}} = (1-\lambda)(\beta\theta_{u,k,w}^{\mathrm{IM}}\theta_{u,k}^{\mathrm{IM}} + (1-\beta)\theta_{u,k,w}^{\mathrm{CM}}\theta_{u,k}^{\mathrm{CM}}) + \lambda P(\omega \mid \phi_k^{\mathrm{TM}}) \tag{14}$$

式中，参数β用来控制朋友模型对用户模型的影响，采用Dirichlet先验平滑，β为：

$$\beta = \frac{|\boldsymbol{M}_u|}{|\boldsymbol{M}_u| + \mu} \tag{15}$$

式中，μ为Dirichlet平滑参数。

2.2 微博质量建模

微博质量是影响用户信息选择的重要因素。本文采取微博长度、附加链接、有无标签、转发数量等指标对微博质量进行建模，具体方法如下。

1) 微博长度：利用标准化指标Length(r)度量微博长度指标，其中$\boldsymbol{R}$是搜索到的所有微博。

$$\text{Length}(r)=\frac{L(r)}{\max\limits_{r'\in \boldsymbol{R}} L(r')} \tag{16}$$

2) 附加链接：用 u 表示微博中链接的数量。

$$\text{Url}(r)=\begin{cases} u & \text{有链接} \\ 0 & \text{否} \end{cases} \tag{17}$$

3) 有无标签：用 h 表示微博中标签的数量。

$$\text{Hashtag}(r)=\begin{cases} h & r>0 \\ 0 & r=0 \end{cases} \tag{18}$$

4) 转发数量：用 $\text{re}(r)$ 表示转发的微博数量。

$$\text{Retweet}(r)=\begin{cases} \dfrac{\text{re}(r)}{\max\limits_{r'\in \boldsymbol{R}} \text{re}(r')} & \text{re}(r')>0 \\ 0 & \text{re}(r')=0 \end{cases} \tag{19}$$

2.3 发布者建模

微博发布者的身份及其领域影响力均会对用户的信息选择产生影响。由于微博发布者在经验、信仰、文笔、价值观等方面的指标很难评判，本文基于主题特征与社交特性构建微博发布者模型。

设$\boldsymbol{A}$为用户查询结果对应的所有作者集合，a代表一个作者。本文基于发布者特性和用户-发布者关系两个维度来衡量微博发布者对用户信息选择行为的影响。

1) 影响力：通常使用粉丝数量及关注人数来衡量。

$$\text{Influence}(a)=\frac{i(a)}{i(a)+o(a)} \tag{20}$$

式中，$i(a)$ 表示用户的粉丝数量；$o(a)$ 表示用户关注人数，影响力越高的用户Influence值越接近于1。

2) 认证名人：将用户是否是认证名人记为 $\text{PC}(a)$。

$$\text{PC}(a)=\begin{cases} c & \text{是} \\ 0 & \text{否} \end{cases} \tag{21}$$

3) 传播能力：用转发率衡量传播能力。

$$\text{Transfer}(a)=\lg(1+c(\text{retweet})) \tag{22}$$

4) 内容关联：发布者社会阅历对说服力影响。

$$\text{Authority}(a,r)=\sum_{k=1}^{K}\omega_T(u,k)\omega_T(r,k) \tag{23}$$

2.4 用户-发布者关系建模

1) 相似度：使用KL-divergence[10]来衡量用户和微博作者在主题分布上的相似度，值越大则说明用户和作者在主题上越有共同点。

$$\text{Similar}(u,a)=1/\text{KL}(\theta_u \,\|\, \theta_a) \tag{24}$$

2) 共同关注：该指标体现了用户与微博作者关注的人的交集。

$$\text{Jaccard}(u,a)=\frac{\left|\text{Followee}(u)\cap \text{Followee}(a)\right|}{\left|\text{Followee}(u)\cup \text{Followee}(a)\right|} \tag{25}$$

2.5 搜索结果排序

按照如下思路进行指标融合，如果用户搜索出的微博是高质量的，该微博的作者具有某些特征；如果用户对该作者感兴趣，则该结果会被排在前面。

$$\omega_{u,a}=\boldsymbol{\alpha}_1^{\mathrm{T}}\begin{pmatrix} \text{Length}(r) \\ \text{Url}(r) \\ \text{Hashtag}(r) \\ \text{Retweet}(r) \end{pmatrix}\boldsymbol{\alpha}_2^{\mathrm{T}}\begin{pmatrix} \text{Follower}(a) \\ \text{Transfer}(a) \\ \text{PC}(a) \\ \text{Authority}(a,r) \end{pmatrix}\boldsymbol{\alpha}_3^{\mathrm{T}}\begin{pmatrix} \text{Similar}(u,a) \\ \text{Jaccard}(u,a) \end{pmatrix} \tag{26}$$

式中，$0\leqslant\omega_{u,a}\leqslant 1$；参数$\boldsymbol{\alpha}_1^{\mathrm{T}}$、$\boldsymbol{\alpha}_2^{\mathrm{T}}$、$\boldsymbol{\alpha}_3^{\mathrm{T}}$用于控制用户各个指标项所占的权重。将作者权重归一化 $\sum\limits_{a\in \boldsymbol{A}}\omega_{u,a}=1$，即可得到每个作者的相对权重。

通过以上部分的探讨，为搜索者建立了个性化的用户模型，并寻找用户与每一条微博作者的关联，提供了用于个性化排序的作者模型。单独使用用户模型会忽略用户对作者的偏好，而只使用作者模型会忽略用户对于微博内容本身的需求。利用语言模型，很好地将用户模型和作者模型结合起来，将作者模型作为一条微博的先验融入语言模型的排序方法中：

$$P(\boldsymbol{D},\boldsymbol{Q},u,a)\propto\left(\sum_{k=1}^{\boldsymbol{K}}P(\boldsymbol{Q}|\hat{\theta}_{u,k,w}^{\text{IM}})P(\boldsymbol{D}|\hat{\theta}_{u,k,w}^{\text{IM}})\right)\omega_{u,a} \tag{27}$$

式中，$P(\boldsymbol{Q}|\hat{\theta}_{u,k,w}^{\text{IM}})=\prod P(\omega|\hat{\theta}_{u,k,w}^{\text{IM}})$；$P(\boldsymbol{Q}|\hat{\theta}_{u,k,w}^{\text{IM}})$ 和 $P(\boldsymbol{D}|\hat{\theta}_{u,k,w}^{\text{IM}})$ 是搜索结果$\boldsymbol{D}$和查询$\boldsymbol{Q}$的主题特定的个性化得分；$\omega_{u,a}$是用户针对该微博作者的先验概率。

该方法将排序分为3部分，首先运用用户模型进行查询的消歧，通过主题模型预测用户拟出查询$\boldsymbol{Q}$时所指的主题；然后运用用户模型计算搜索结果$\boldsymbol{D}$所属的主题；再运用作者模型预测用户可能对微博作者的兴趣，作为先验概率融入排序过程。

$P(\boldsymbol{Q}\,|\,\hat{\theta}_{u,k,w}^{\text{IM}})$ 通过每个词概率相乘得到，由于微博的短文本特性将其标准化：

$$P(\boldsymbol{Q}\,|\,\hat{\theta}_{u,k,w}^{\text{IM}})=(\prod P(\omega\,|\,\hat{\theta}_{u,k,w}^{\text{IM}}))^{\frac{1}{n(\omega)}} \tag{28}$$

3 实验评价

3.1 数据集

为构建主题模型，本文通过爬虫抓取了新浪微博的数据。随机选取了5 138个用户，259万条微

博。删除了“僵尸”用户，剩余用户5 003个。对少于10字的微博进行了过滤，剩余212万条微博。基于该数据集，利用Mallet训练主题模型[17]。

为检验个性化搜索方法的有效性，本文采用了用户参与评分方法。实验共选择了33位活跃用户(半年发微博数量多于200)。为了构建用户个性化兴趣模型，抽取每位用户半年内的微博，并抓取了用户的朋友列表，通过2.1节所述方法计算用户和朋友的关系，选取关系最近的前10位朋友，抓取这些朋友的微博，以及朋友的粉丝数等信息。图1展示了本次实验的完整流程。

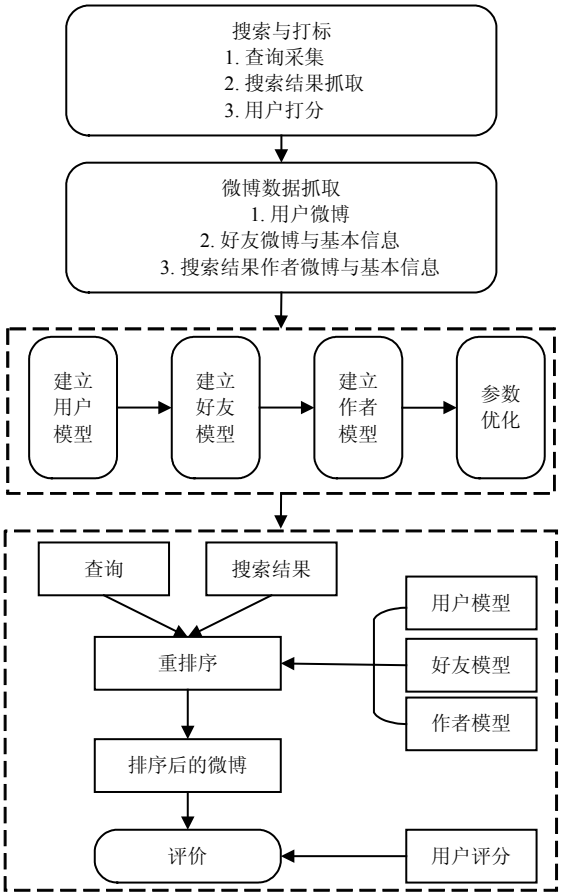

图1 实验流程图

参照文献[6]的方法，将查询关键词分为4种类型，如表3所示，实验要求每个用户提交多于4个关键词，且需至少包含3种类型[19]。

由于用户在进行微博搜索时通常只关注前两页的返回结果，因此针对每个关键词抓取微博返回的前两页搜索结果(20条)进行个性化排序。为了度量微博发布者的特征，实验抓取了返回结果的作者在半年内发布的微博、每条微博的转发次数与评论次数、发布者的粉丝数、关注数、关注列表和认证信息等。实验共收集查询139个，将搜索结果不足20个的去除掉，最终收集到共125组搜索结果。由于搜索结果中往往会出现僵尸用户所发信息，参考文献[20]的僵尸用户识别方法，将关注小于10，粉丝大于100小于200、用户名使用人名和动作的组合、发布微博含有大量的非正常符号等特征的僵尸用户及所发微博过滤，并将搜索结果交由用户评价。将搜索结果随机排序，如果该条结果是用户想要的，用户打1分，否则打0分。

表3 查询类型分类

查询类型	比例/%	例子
时事	29.6	澳大利亚VS韩国、姜辉 马航、淘宝 工商总局
长期兴趣	32.8	汉服、古风、cosplay、机器学习、合肥 美食
实体	48.8	方滨兴、大学勤工助学中心、邓超、科比、天猫
模糊词	24.8	cos(余弦、动漫角色扮演)、小马甲(服饰、新浪大V回忆专用小马甲)、一步之遥(姜文电影、一种俗语)、兰州兰州(低苦艾的一首歌、地名)、小米(粗粮、热门手机品牌)、金球奖(美国的一个电影与电视奖项、国际足球联合会金球奖)

搜索查询平均长度为1.26，搜索结果平均长度为18.94。用户评价相关的搜索结果占比13.24%。搜索结果的作者中，认证微博用户共占比6.40%。搜索结果中含标签(#Hashtag)的占比8.21%，含链接的占比2.45%。

3.2 基准模型

为了验证模型的有效性，对查询似然模型(B-QM)和协同个性化搜索主题−语言模型(B-CM)进行了程序实现。B-QM是语言模型的经典方法，可以对搜索关键词与微博内容的相关性进行度量。B-CM是由文献[6]提出的个性化微博搜索方法，该方法运用主题模型与语言模型进行个性化搜索。此外，实验将本文方法逐块拆分为A-AMQ模型(仅考虑微博质量)、A-AMQF模型(考虑微博质量与作者特征)、以及A-AMQFA模型(考虑微博质量、作者特征以及用户-发布者关系)。并将上述模型与本文A-AMQFA的个性化搜索结果进行对比。排序结果使用P@N(前N个结果的正确率)和MAP(宏平均正确率)指标进行评价。

3.3 实验结果

为了对本文模型进行训练，对相关参数的设置如表4所示。表中参数值是针对本次实验所用数据，经过多次实验寻优的结果。其他数据所对应的最优参数值将会有所变化。

表5、表6分别给出了本文方法与基准方法的效果对比。可以看出，与QM和CM方法相比，本文方

法具有更好的个性化搜索精度。而且，随着微博质量、发布者特征和用户−发布者关系等信息的逐步加入，个性化搜索的效果不断改进。这表明，本文方法构建了主题和语言的双层模型，将微博质量、发布者特征和用户−发布者关系融入个性化搜索模型，对提高微博个性化搜索效果具有较好的作用。

表4　模型参数

参数	描述	值
λ	用户模型中JM平滑的参数	0.2
β	朋友模型中Dirichlet平滑的参数	80
c	作者模型中作者权威度中认证名人的参数	0.2
u	作者模型中微博质量中含链接的参数	0.25
h	作者模型中微博质量中含标签的参数	0.25
$\boldsymbol{\alpha}_1^{\mathrm{T}}$	作者模型中用于控制微博质量的参数	(0.4,0.2,0.2,0.2)
$\boldsymbol{\alpha}_2^{\mathrm{T}}$	作者模型中用于控制作者特征的参数	(0.3,0.2,0.1,0.4)
$\boldsymbol{\alpha}_3^{\mathrm{T}}$	作者模型中用于控制用户与作者关联的参数	(0.55,0.45)

表5　用户−发布者模型与基准模型在MAP指标下的表现

模型	MAP(3)	MAP(6)	MAP(9)	MAP(12)	MAP(15)
QM	0.091 067	0.096 896	0.097 905	0.098 472	0.099 776
CM	0.105 067	0.108 798	0.108 221	0.108 689	0.108 841
AMQ	0.125 067	0.126 484	0.124 264	0.122 397	0.121 052
AMQF	0.124 667	0.127 685	0.126 504	0.125 561	0.124 194
AMQFA	0.126 933	0.129 347	0.126 268	0.124 903	0.125 218

表6　用户−发布者模型与基准模型在P@*N*指标下的表现

模型	P@5	P@10	P@15
QM	0.076 16	0.089 28	0.097 600
CM	0.086 72	0.098 24	0.101 867
AMQ	0.104 96	0.110 08	0.106 773
AMQF	0.109 12	0.110 40	0.106 667
AMQFA	0.110 72	0.112 48	0.108 373

3.4　参数敏感性分析

本文利用微博质量、发布者特征和用户−发布者关系三方面信息对主题语言双层模型的排序结果进行修正。由于上述三方面信息均由多个指标进行度量，为了验证各度量指标的影响，本文对相关指标进行敏感性实验。

微博质量由微博长度(*L*)、链接(*U*)、标签(*H*)、转发(*R*)4个指标组成。将不同指标进行组合，研究基于不同微博质量度量指标的情况下，个性化搜索结果的变化。在作者模型中仅考虑微博质量对结果进行验证，如图2所示，文档长度、链接与标签对结果排序最为明显。因此本文为含有链接、标签的长微博赋予了更高的权重，将可能含有更多话题与标签的微博呈现给用户。而转发数指标并没有太好的效果，这表明用户在搜索时更关注的是微博的内容，而不是该微博是否热门。本文对参数进行了优化，在此基础上融合微博作者的特征做了进一步的因素分析。

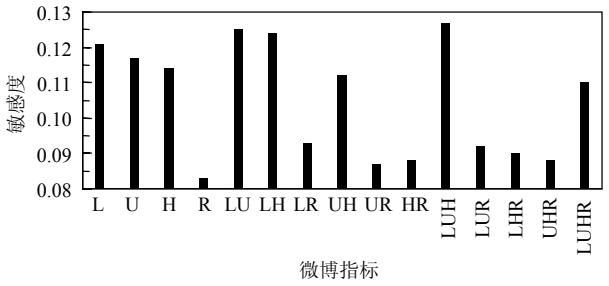

图2　微博质量各特征敏感性

发布者特征包含4个要素，即作者影响力(I)、微博的传播能力(T)、是否是认证用户(P)、在主题上与该微博的关联(A)。本文在融合微博质量因素的作者模型的基础上进一步对这4个要素进行衡量。如图3所示，微博与作者在主题上的关联、作者的影响力以及传播能力对模型影响最为明显。这一结果与实际相符，人们在关注微博的时候，往往会把作者与该微博的主题联系在一起，会在意该作者是否在该主题上专业、有影响力；而影响力与传播能力也直观地表现了这一特征。而用户对传播能力并不在意，说明用户并不关注该作者的微博是否总是成为热门微博。本文进行参数优化后，进一步地融合用户对作者的关联。

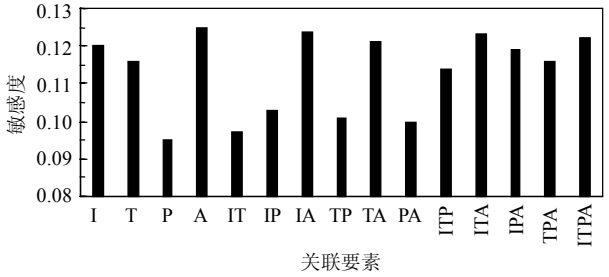

图3　用户与作者关联各特征敏感性

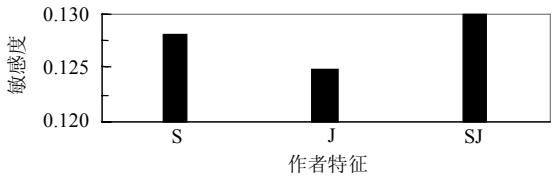

图4　作者特征各特征敏感性

本文采用了作者与用户的主题相似度(S)与共同关注的人(J)两个指标对用户与作者的关联进行衡量。如图4所示，作者与用户关注的主题相似度更

明显，说明用户更愿意关注有共同兴趣的人，而共同关注特征并不明显，可能是因为用户关注的都是人尽皆知的公共用户，不能体现用户在所感兴趣主题方面的特征。

4 结 束 语

该文提出了考虑用户-发布者关系建模的个性化微博搜索模型，该模型在考虑用户对微博兴趣的同时，也有效利用了用户的多种社交因素来发现用户对于主题的兴趣，并在此基础上提出了考虑用户-发布者关系的个性化搜索方法。通过从各个角度对微博作者的刻画，将用户对微博以及微博作者的兴趣有效展现了出来。最后进行了基于真实用户集的个性化搜索实验，并对模型的各个部分进行了分析与敏感性实验。实验表明，本文的模型效果可有效对微博搜索结果进行个性化排序。

由于数据集的限制，本文无法获取用户与微博作者的显性关联特征。下一步将获取更多的微博数据集，寻找用户与微博作者的显性关联，对微博作者在主题上对用户的影响(类似于Twitter Rank)进行分析。此外，该模型并没有考虑用户在主题上的兴趣随着时间推移而变弱的情况，将在后续研究中进一步讨论。

本文研究工作得到了佛山市科技创新项目(2016AG100792)的资助，在此表示感谢！

参 考 文 献

[1] 中国互联网信息中心. 第38次中国互联网发展状况统计报告 [EB/OL]. [2016-08-07]. http://www.cnnic.cn/hlwfzyj/hlwxzbg/hlwtjbg/ 201702/0201702035488526110 92.pdf.
CNNIC. 38th statistical report on the development of Internet Network in China[EB/OL]. [2016-08-07]. http://www.cnnic.cn/hlwfzyj/hlwxzbg/hlwtjbg/201702/ 0201702035488526110 92.pdf.

[2] CHEN K, CHEN T, ZHENG G, et al. Collaborative personalized tweet recommendation[C]//35th International ACM SIGIR Conference on Research & Development in Information Retrieval. New York: ACM, 2012: 661-670.

[3] TAIKI M, KAZUHIRO S. Improving pseudo-relevance feedback via micro-document selection[J]. IPSJ Journal, 2014, 55: 1585-1594.

[4] 卫冰洁, 王斌. 面向微博搜索的时间感知的混合语言模型[J]. 计算机学报, 2014, 37(1): 229-237.
WEI Bing-Jie, WANG Bin. Time-aware mixed language model for microblog search[J]. Chinese Journal of Computers, 2014, 37(1): 229-237.

[5] MASSOUDI K, TSAGKIAS M, RIJKE M D, et al. Incorporating query expansion and quality indicators in searching microblog posts[J]. Lecture Notes in Computer Science, 2011, 66(11): 362-367.

[6] JAN V, KENNETH W T. Collaborative personalized twitter search with topic-language models[C]//37th International ACM SIGIR Conference on Research & Development in Information Retrieval. New York: ACM, 2014: 53-62.

[7] HARVEY M, CRESTANI F, CARMAN M J. Building user profiles from topic models for personalized search[C]// Proceedings of the 22nd ACM International Conference on Information & Knowledge Management. [S.l.]: ACM, 2013: 2309-2314.

[8] ZHAI C. Statistical language models for information retrieval[M]. New York: Morgan and Claypool Publishers, 2008.

[9] MANNING C D, RAGHAVAN P, SCHÜTZE H. Introduction to information retrieval[M]. Cambridge: Cambridge University Press, 2008.

[10] TEEVAN J, RAMAGE D. Twitter search: a comparison of micro blog search and web search[C]//ACM International Conference on Web Search & Data Mining. Hong Kong, China: ACM, 2011: 35-44.

[11] 徐雅斌, 石伟杰. 微博用户推荐模型的研究[J]. 电子科技大学学报, 2015, 44(2): 254-259.
XU Ya-bin, SHI Wei-jie. Research on microblog user recommendation model[J]. Journal of University of Electronic Science and Technology of China, 2015, 44(2): 254-259.

[12] HANNON J, BENNETT M. Recommending Twitter users to follow using content and collaborative filtering approaches[C]//Proceedings of the ACM Conference on Recommender Systems. New York: ACM, 2010: 199-206.

[13] SHANG Y. A new interest-sensitive and network-sensitive method for user recommendation[C]//IEEE Eighth International Conference on Networking, Architecture and Storage(NAS). Washington: IEEE Computer Society, 2013: 242-246.

[14] LOKHOV A Y, MEZARD M, OHTA H. Inferring the origin of an epidemic with dynamic message-passing algorithm[J]. Phys Rev E, 2014, 90(1): 012801.

[15] PRAKASH B A, VREEKEN J, FALOUTSOS C. Spotting culprits in epidemics: How many and which ones?[C]//IEEE International Conference on Data Mining. Brussels: IEEE Computer Society, 2012, 12: 11-20.

[16] ZHU K, YING L. Information source detection in the SIR model: a sample path based approach[J]. IEEE/ACM Transactions on Networking, 2013, 24(1): 408-421.

[17] 张永棠. 基于代换加密的隐私保护协同过滤推荐算法[J]. 新疆大学学报(自然科学版), 2017, 34(4): 446-451.
ZHANG Yong-tang. An algorithm of cooperative filtering for privacy protection based on substitution encryption[J]. Journal of Xinjiang University (Natural Science Edition), 2017, 34(4): 446-451.

编 辑 蒋 晓