基于维基百科类别图的推特用户兴趣挖掘

刘小捷[1] 吕晓强[1] 王晓玲[1] 张 伟[1] 赵 安[2]
(华东师范大学上海市高可信计算重点实验室 上海 200062)[1]
(中国科学院电子学研究所苏州研究院 江苏 苏州 215123)[2]

摘 要 以Twitter为代表的社交网络在人们的生活中发挥着重要作用,其庞大的用户群体给社交网络数据挖掘带来了巨大的价值。社交网络用户兴趣建模方法被广泛研究,并被用于提供个性化推荐。文中提出了一种基于维基百科类别图的Twitter用户兴趣挖掘和表示方法。首先,该方法根据用户活跃度的差异,分别采用基于推文内容的方法和基于关注账号信息的方法来实现活跃用户与非活跃用户的兴趣挖掘。然后,在维基百科类别图上使用个性化PageRank算法进一步拓展用户兴趣,生成维基百科类别表示的用户兴趣画像。在推文推荐的应用背景下,对用户兴趣建模策略进行了实验分析和比较。实验结果表明,与现有的Twitter用户兴趣挖掘方法相比,所提方法显著提升了推文推荐效果,能够有效地改进用户兴趣挖掘效果。

关键词 社交网络,用户兴趣,个性化PageRank,推文推荐
中图法分类号 G633.67 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2019.09.010

Mining User Interests on Twitter Using Wikipedia Category Graph

LIU Xiao-jie[1] LV Xiao-qiang[1] WANG Xiao-ling[1] ZHANG Wei[1] ZHAO An[2]
(Shanghai Key Laboratory of Trustworthy Computing, East China Normal University, Shanghai 200062, China)[1]
(Institute of Electronics, Chinese Academy of Sciences, Suzhou, Jiangsu 215123, China)[2]

Abstract Social network such as Twitter plays an important role in life, and the huge number of users makes social network data mining valuable. User interest modeling on social networks has been studied widely, and is used to provide personalized recommendations. This paper proposed a novel user interest mining and representation approach based on Wikipedia Category Graph. User interest profile is represented as a wikipedia category vector. First, according to the degree of user's activeness, an interest mining method based on tweets is proposed for active users, and another method based on names and descriptions of followees is proposed for passive users. Then, user interest is extended and generalized based on Wikipedia Category Graph by personalized PageRank algorithm, and user interest profile is represented by wikipedia categories. The proposed interest modeling strategy was evaluated in the context of a tweet recommendation system. The results shows that the proposed approach improves the quality of recommendation significantly compared with the state-of-the-art Twitter user interest modeling approachs, which means it can provide a more effective user interest profile.

Keywords Social network, User interest, Personalized PageRank, Tweets recommendation

1 引言

近年来,随着移动互联网的发展,全球社交网络的规模迅猛增长。社交网络在人们的日常活动中发挥着越来越重要的作用,越来越多的人热衷于在社交媒体上关注和评论自己感兴趣的事物,分享生活中的点点滴滴。社交媒体平台拥有庞大的用户群体,其中Twitter和Facebook的月活跃用户数已经分别达到3.28亿和20.1亿[1)]。庞大的用户群体在社交媒体平台上生成了海量的数据信息,社交媒体数据挖掘带来了巨大的价值,因此社交媒体数据挖掘成为了当前的研究热点。

社交媒体平台上的海量数据给用户带来了一个严重的问题——信息过载,这一问题造成用户的个性化需求不断增加。个性化的主要问题之一就是建立用户的个人画像。用户画像的构建是获取、提取和表示用户特征的过程[1]。社交网络领

1) http://epaper.21jingji.com/html/2017-07/28/content_67268.htm

到稿日期:2018-07-02 返修日期:2018-09-13 本文受国家自然科学基金(61472141),国家重点研发计划(2017YFC0803700),上海市重点学科建设项目(B412),上海市可信物联网软件协同创新中心(ZF1213)资助。

刘小捷(1994—),男,硕士,主要研究领域为社交媒体数据挖掘;**吕晓强**(1993—),男,硕士,主要研究领域为数据挖掘;**王晓玲**(1975—),女,教授,博士生导师,CCF会员,主要研究领域为数据分析与数据管理,E-mail:xlwang@sei.ecnu.edu.cn(通信作者);**张 伟**(1988—),男,博士,副研究员,主要研究领域为数据挖掘与自然语言处理;**赵 安**(1992—),女,硕士,主要研究领域为自然语言处理与图像处理。

域中的用户画像可以表示为每个用户各种类型的相关信息，这些信息可能是年龄、性别、国家等基本信息或者是代表其兴趣的关键字。从海量而杂乱的社交媒体数据中精确地挖掘用户的特征，构建用户画像，对于广告投放、精准营销、推荐系统等个性化服务具有巨大的商业价值。

本文以 Twitter 平台为例，深入研究用户兴趣画像构建的问题，利用社交网络数据挖掘用户的潜在兴趣。用户兴趣建模通常分为两类：基于文本内容和基于用户行为[2]。本文的研究是基于文本内容的用户兴趣建模。本文的主要贡献如下：

(1)根据用户活跃度的差异，通过采用不同种类的社交网络文本数据，分别提出了适用于活跃用户和非活跃用户的兴趣挖掘算法，生成由维基百科类别表示的用户兴趣。

(2)根据维基百科类别与原始文本内容之间的语义相似度，设计了一种新的权重分配策略，充分考虑了用户对不同类别的实际兴趣倾向。

(3)在原始兴趣的基础上，根据原始兴趣在维基百科类别图(Wikipedia Category Graph，WCG)中的结构关系，提出了一种基于个性化 PageRank 算法的用户兴趣扩展方法。

(4)在推文推荐的应用背景下，对用户兴趣建模策略进行了实验分析和比较，结果表明，本文提出的用户兴趣画像构建方法在所有评估指标上的表现都是最好的。

2 相关工作

通过文本内容进行用户兴趣挖掘的传统方法主要是基于主题模型。由于推文的短文本特性，标准的 LDA 主题模型并不适用。为了解决这一问题，一些研究将用户的所有推文集合作为一个文档，如 TwitterRank[3]。事实上，这种处理可以被认为是 Author-Topic 模型[4]在推文上的应用，因为每个文档都有一个作者。另外，基于一条推文仅由一个主题构成假设，Zhao 等提出了 Twitter-LDA[5]。除用户自身发布的推文外，一些研究还利用其他类型的用户信息。例如，Chen 等[6]使用关注账号的推文来发现用户的兴趣主题。相反，Hannon 等[7]使用粉丝和粉丝的推文来扩展用户的属性，并使用 TF-IDF 对提取的关键字进行加权。

由于推特的短文本特性，一条推文中包含的信息量有限。近年来，越来越多的方法使用其他数据源对原始推文进行语义丰富。这些方法通常将文本内容中提到的术语链接到知识库(如 Wikipedia)中的实体概念，并创建基于实体概念的用户兴趣。这些知识库由于包含了概念及其关系，因此为推断文本内容的潜在语义提供了技术手段。例如，Lu 等[8]从用户推文中提取实体概念，通过随机游走的方式在知识图谱上找到相关实体概念进行扩展。Michelson 等[9]首先从用户推文中提取一组维基实体，然后通过遍历和分析提取的实体所属的维基百科类别来识别高层次的用户兴趣。Siehndel 等[10]提出了 TwikiMe，通过从用户的推文中提取实体并将其链接到维基百科的 23 个顶级类别来生成用户兴趣画像。Kapanipathi 等[11]使用维基百科来发现推文中的实体，将原始兴趣映射到由 WCG 转换而成的层次结构中，并通过传播激活的方式来推断用户兴趣。

针对推文数据稀疏的问题，一些研究也使用了其他类型的用户信息进行语义丰富，用于推断非活跃用户的兴趣。例如，Lim 等[12]通过使用维基百科类别将用户关注的名人划分为不同的兴趣类别，然后根据用户对不同兴趣类别的名人关注数量来确定用户的相对兴趣。Besel 等[13]和 Faralli 等[14]将关注账号的名称与维基百科实体相关联，然后使用这些实体概念信息来推断用户兴趣画像。Piao 等[15]提出从关注账号的描述信息中提取实体，并将实体关联至 DBpedia 知识库从而生成用户兴趣画像。

现有的相关研究仅考虑了单一类型的用户，无法同时适用于活跃用户和非活跃用户。另外，这些方法在用户兴趣的集成过程中没有充分利用维基百科类别之间的结构关系。因此，本文针对上述不足进行了改进，提出了一种能够同时适用于活跃用户和非活跃用户的用户兴趣画像构建方法，并在维基百科类别图上使用个性化 PageRank 算法来进一步拓展用户兴趣。

3 相关工作

本节主要介绍用户兴趣画像模型的构建方法，包括原始兴趣挖掘和兴趣拓展两部分。已有的研究主要集中于经常主动发布推文的活跃用户，通过分析用户的推文生成相应的用户兴趣画像。然而，Twitter 中也存在大量的非活跃用户，其很少甚至从来不发布推文。因此，本文针对 Twitter 中活跃和非活跃两种类型的用户，使用不同的文本数据进行用户的原始兴趣挖掘；此外，本节将详细介绍用户兴趣的拓展方法。

本文提出的 Twitter 用户兴趣画像构建流程如图 1 所示。图 1 中，Twitter 用户兴趣画像构建由 5 部分组成：1)用户文本内容，包括用户发布的推文、用户关注账号的名称和个人描述信息，这些内容使用 Twitter API 进行采集；2)用户原始兴趣挖掘，通过实体抽取和实体链接技术处理文本内容，获取维基百科类别并进行权重分配；3)用户原始兴趣，参考定义 2 和定义 4，生成活跃用户和非活跃用户的原始兴趣；4)兴趣拓展，在维基百科类别图上使用个性化 PageRank 算法拓展用户兴趣；5)用户兴趣画像，参考定义 1，生成维基百科类别表示的用户兴趣画像。

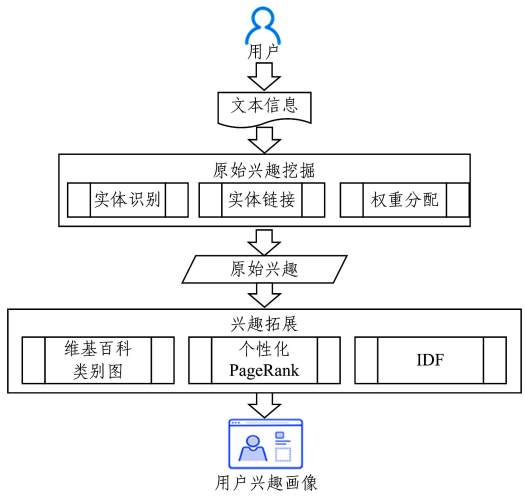

图 1 用户兴趣画像构建流程

Fig. 1 Process of user interest profile construction

定义 1(用户兴趣画像) 用户 u 的兴趣画像 P_u 是一组

加权的用户兴趣（维基百科类别）。每个兴趣 $i\in I$ 的权重 $w(u,i)$ 表示兴趣 i 对用户 u 的重要程度。其中，$u\in U$，$i\in I$，U 表示用户集合，I 表示用户兴趣集合。

$$P_u=\{(i,w(i,u))\mid i\in I\} \tag{1}$$

3.1　原始兴趣挖掘

将文本内容中直接提取的实体概念所属的维基百科类别称为原始兴趣，可以基于知识库进一步增强这些原始兴趣。在原始兴趣挖掘过程中，首先通过命名实体识别技术抽取文本的实体。然后通过实体链接技术将文中实体关联至维基百科中对应的实体概念，并获取实体所属的维基百科类别。由于实体链接工具 TagMe[1)] 在处理短文本时具有非常好的性能，已经在相关研究中被广泛应用，因此本文同样使用 TagMe。最后，根据相应的权重分配策略计算维基百科类别的权重，生成用户的原始兴趣。

维基百科中的一个实体概念属于多个维基百科类别，而在现有的相关研究工作中，对同一实体所属的维基类别的权重的取值相同，未考虑用户对不同类别的实际兴趣倾向。例如，推文“Trump's tax cuts didn't benefit American workers”中，“Trump”所对应的维基百科实体概念 Donald Trump[2)] 属于 Presidents of the United States，American billionaires，American people of German descent，1946 births，Living people 等 52 个维基百科类别。显然，这一推文反映出了用户对 Presidents of the United States 这一类别的兴趣程度高于 American people of German descent，1946 births，Living people 等类别。因此，本文根据维基百科类别与原始文本内容之间的语义相似度设计了一种新的权重分配策略。

针对短文本语义相似度已有相关研究，如 Kenter 等[16] 利用词向量进行计算。由于短文本中所包含的单词数量较少，计算短文本语义相似度的常用方法是将文本中出现的单词的词向量取平均得到文本的向量表示，再计算文本向量间的余弦相似度。其中，每个单词的词向量可以通过 Word2Vec[17] 使用浅层神经网络语言模型学习得到。该方法已经被证明在文本的分类和聚类中是有用的，因为词向量均值表示了文本的整体主题。本文采用该方法计算短文本的相似度，短文本 s_1 和 s_2 之间的相似度计算公式为：

$$sim(s_1,s_2)=cosine(\boldsymbol{Vec}_{s_1},\boldsymbol{Vec}_{s_2}) \tag{2}$$

其中，$\boldsymbol{Vec}_s$ 为短文本 s 的向量表示，计算公式如下：

$$\boldsymbol{Vec}_s=\frac{1}{n}\sum_{w\in s} f_w * v_w \tag{3}$$

其中，n 为 s 中的单词总数，f_w 和 $\boldsymbol{v}_w$ 分别为 s 中单词的词频和词向量。

根据本文提出的基于文本相似度的权重分配策略，给定短文本和实体概念，维基百科类别的权重计算公式的定义如下：

$$w_e(s,c)=\begin{cases}2^{-rank(sim(s,c))}, & 若\ c\in C_e\\ 0, & 若\ c\notin C_e\end{cases} \tag{4}$$

其中，C_e 为实体概念 e 所属的维基百科类别集合，$rank(sim(s,c))$ 表示根据 $sim(s,c)$ 得到的 c 在 C_e 中的排序值，$sim(s,c)$ 表示根据式(2)得到的 s 和 c 之间的语义相似度。

与之前的一些研究[6,18-20] 类似，本文认为如果用户至少发布了 100 条推文，则该用户处于活跃状态。根据用户发布的推文数量，将 Twitter 用户分为活跃和非活跃两类。对于不同类型的用户，使用不同种类的文本内容进行用户的原始兴趣挖掘。

(1)活跃用户原始兴趣挖掘

对于活跃用户，本文使用用户发布的推文进行挖掘。根据定义 2 获取活跃用户的原始兴趣，活跃用户的原始兴趣挖掘过程如算法 1 所示。

算法 1　活跃用户原始兴趣挖掘算法 AUTI

输入：推文集合下

输出：用户原始兴趣 TI

```
1. initial TI=∅;
2. for each t∈T do
3.   E_t=NER(t);//获取 t 中的实体集合
4.   for each e∈E_t do
5.     C_e=NEL(e);//获取 e 所属类别集合 C_e
6.     for each c∈C_e do
7.       w_c=w_e(t,c);
8.       if c in TI then
9.         TI(c).add(w_c);
10.      else
11.        TI.set(c,w_c);
12.      end if;
13.    end for;
14.  end for;
15. end for;
16. totalWeight=Σ_{c in TI} w_c;
17. for c in TI do
18.  TI.update(c,w_c/totalWeight);
19. end for;
```

定义 2(活跃用户原始兴趣)　计算用户推文中每个实体概念所属的维基百科类别的权重，得到由维基百科类别构成的兴趣，并将其作为活跃用户的原始兴趣，定义如下：

$$TI=\{(c_1,w_{c_1}),(c_2,w_{c_2}),\cdots,(c_n,w_{c_n})\} \tag{5}$$

其中，w_{c_i} 为原始兴趣 c_i 的权重，计算公式为：

$$w_{c_i}=\frac{\sum_{t\in T}\sum_{e\in E_t} w_e(t,c_i)}{\sum_{t\in T}\sum_{e\in E_t}\sum_{c\in C_e} w_e(t,c)} \tag{6}$$

其中，T 为用户的推文集合；t 表示其中一条推文($t\in T$)；E_t 为推文 t 中的实体概念集合；e 表示其中一个实体概念($e\in E_t$)；C_e 为 e 所属的维基百科类别集合；c 表示其中一个维基百科类别($c\in C_e$)；$w_e(t,c)$ 表示对于给定推文 t 和实体概念 e，根据式(4)得到的维基百科类别 c 的权重。

(2)非活跃用户的原始兴趣挖掘

对于非活跃用户，由于用户发布的推文数量较少，推文中包含的信息量较少，本文使用用户关注账号的名称和个人描述信息进行挖掘，根据定义 4 获取非活跃用户的原始兴趣。非活跃用户的原始兴趣挖掘过程如算法 2 所示。

算法 2　非活跃用户的原始兴趣挖掘算法 PUFI

输入：用户关注的账号集合 F

1) https://tagme.d4science.org/tagme/

2) https://en.wikipedia.org/wiki/Donald_Trump

输出：用户原始兴趣 FI

```
initial FI=∅;
for each f∈F do
  if f 是名人类账号 then
    E_f={e_{f_name}};
  else
    E_f=NER(f_d); //获取 f_d 中的实体集合
  end if;
  for each e∈E_f do
    C_e=NEL(e); //获取 e 所属类别集合 C_e
      for each c∈C_e do
        w_c=w_e(f_d,c);
        if c in FI then
          FI(c).add(w_c)
      else
          FI.set(c,w_c);
      end if;
    end for;
  end for;
end for;
totalWeight=Σ_{c in FI} w_c;
for c in FI do
  FI.update(c,w_c/totalWeight)
end for;
```

定义 3(名人类账号) 若推特账号为认证账号，且该账号的名称能够链接到维基百科中的实体概念，则将该账号定义为名人类账号。名人类账号不仅是知名人物，还可能是机构、组织等。

定义 4(非活跃用户原始兴趣) 计算用户关注账号信息中每个实体概念所属的维基百科类别的权重，得到由维基百科类别构成的兴趣，并将其作为非活跃用户的原始兴趣，定义如下：

$$FI=\{(c_1,w_{c_1}),(c_2,w_{c_2}),\cdots,(c_n,w_{c_n})\} \tag{7}$$

其中，w_{c_i} 为原始兴趣 c_i 的的权重，计算公式为：

$$w_{c_i}=\frac{\sum_{f\in F}\sum_{e\in E_f} w_e(f_d,c_i)}{\sum_{f\in F}\sum_{e\in E_f}\sum_{c\in C_e} w_e(f_d,c)} \tag{8}$$

其中，F 为用户关注的账号集合；f 表示其中一个关注账号($f\in F$)；f_{name} 表示 f 的名称；f_d 表示 f 的个人描述信息；E_f 为 f 中的实体概念集合，参考定义 3 将 f 分为名人类账号和非名人类账号；若 f 为名人类账号则 $E_f=\{e_{f_{\text{name}}}\}$，$e_{f_{\text{name}}}$ 表示 f_{name} 对应的实体概念；若 f 为非名人类账号则 E_f 为 f_d 中的实体概念集合。e 表示其中一个实体概念($e\in E_f$)；C_e 为 e 所属的维基百科类别集合；c 表示其中一个维基百科类别($c\in C_e$)；$w_e(f_d,c)$表示对于给定个人描述信息 f_d 和实体概念 e，根据式(4)得到的维基百科类别 c 的权重。

3.2 兴趣拓展

由于原始兴趣中的维基百科类别是实体概念直接所属的类别，大多较为具体，因此，本文在原始兴趣的基础上提出了一种基于 WCG 的用户兴趣拓展方法。

由于 WCG 中包含了一些与兴趣主题无关的类别，如维基百科自身的管理类别[1)]。因此，需要对完整的 WCG 进行预处理，去除无关类别。本文仅保留了 Main topic classifications[2)] 所包含的类别，其中包括 Arts，Politics，Sports 等宽泛的类别，以及 Los Angeles Lakers coaches，German military leaders of World War II 等更为具体的类别。

在用户兴趣拓展过程中，先使用原始兴趣中的维基百科类别在 WCG 中向上层遍历，查找所有祖先类别，构成基于维基百科类别的用户兴趣图，由类别作为兴趣节点，节点之间的边表示兴趣之间的关系。再使用个性化 PageRank 算法[21]在用户兴趣图上进行随机游走，迭代更新兴趣图中节点的权重，直至达到平稳分布。用户兴趣图中，兴趣节点在第 $t+1$ 次迭代过程中的权重更新公式为：

$$PR(c_i)_{t+1}=d\sum_{c_j\in Sub(c_i)}\frac{PR(c_i)_t}{N(c_j)}+(1-d)w_{c_i} \tag{9}$$

其中，c_i 为维基百科类别表示的兴趣节点，$PR(c_i)_t$ 代表兴趣节点 c_i 第 t 次迭代后的权重，$Sub(c_i)$表示 c_i 的子节点集合，$N(c_j)$表示 c_j 的父节点数量，w_{c_i} 为 c_i 在原始兴趣(TI 或 FI)中的权重，d 为阻尼系数。则在第 $t+1$ 次迭代过程中，根据式(9)可以得到所有兴趣节点权重构成的向量更新公式：

$$\boldsymbol{R}_{t+1}=d\boldsymbol{M}\boldsymbol{R}_t+(1-d)\boldsymbol{I} \tag{10}$$

其中，$\boldsymbol{R}_t$ 为第 t 次迭代过程中兴趣节点权重构成的向量，$\boldsymbol{I}$ 为兴趣节点在用户原始兴趣(TI 或 FI)中的权重构成的个性化向量，矩阵 $\boldsymbol{M}$ 的定义如下：

$$\boldsymbol{M}_{ij}=\begin{cases}\dfrac{1}{N(c_j)}, & \text{若 } c_j \text{ 为 } c_i \text{ 的子节点}\\ 0, & \text{其他}\end{cases} \tag{11}$$

经过多次迭代收敛后得到兴趣向量 $\boldsymbol{R}$。

最后，为了体现用户之间的兴趣差异性，使用 IDF(Inverse Document Frequency)更新兴趣向量 $\boldsymbol{R}$ 中维基百科类别 c 的权重 $PR(c)$，定义如下：

$$PR(c)'=PR(c)\times\lg\frac{|U|}{UF(c)+1} \tag{12}$$

其中，U 为用户集合，$UF(c)$为用户兴趣包含的用户数量。用户兴趣图中的各个维基百科类别表示拓展后的用户兴趣，更新后的权重表示用户的感兴趣程度，参考定义 1 生成用户的兴趣画像。

4 相关工作

4.1 实验数据

在数据集的构造过程中，首先通过 Twitter Stream API[3)]，随机选择 1000 位 Twitter 用户。采集这些用户的推文、用户关注账号的名称和个人描述。另外，采集了用户最近点赞的 5 条至少包含一个维基百科实体概念的推文，满足这一条件的用户共 828 位，其中包括 801 位活跃用户和 27 位非活跃用户。

对于活跃用户的兴趣画像构建，使用 801 位活跃用户最近发布的 100 条推文进行用户兴趣挖掘。对于非活跃用户的

1) https://en.wikipedia.org/wiki/Category:Wikipedia_administration

2) https://en.wikipedia.org/wiki/Category:Main_topic_classifications

3) https://developer.twitter.com/en/docs/tweets/sample-realtime/api-reference/decahose

兴趣画像构建，由于非活跃用户的数量较少，本文与 Piao 等的方法相同[15]，使用全部用户（828 位）的关注账号信息进行用户兴趣挖掘，而不考虑用户的推文信息。

为了评价用户兴趣画像模型的构建质量，我们对不同用户兴趣建模策略在推文推荐系统中的推荐效果进行了评估。活跃用户的推荐推文候选集由 801 位用户最近点赞的 5 条推文构成，共 3882 条不同的推文。非活跃用户的推荐推文候选集由 828 位用户最近点赞的 5 条推文构成，共 4015 条不同的推文。

4.2　评估方法

本文的主要目标是在推文推荐的背景下分析和比较不同的用户兴趣建模策略。将不同用户兴趣建模策略生成的兴趣画像作为输入，比较相同推荐算法下所实现的推荐效果，而不是旨在优化推荐效果。与先前一些研究[15,18-19,22]中的方式相同，本文应用了一种轻量级的基于内容的推荐算法来生成推荐。

给定通过相同兴趣建模策略生成的用户兴趣画像 P_u 及候选推文集合 $M=\{P_{t1},P_{t2},\cdots,P_{tm}\}$，推荐算法根据 P_u 与 $P_{ti}(i=1,2,\cdots,m)$ 中由各个兴趣的权重构成的兴趣向量之间的余弦相似度对候选推文进行排序，然后给出相似度得分最高的前 N 条推文。

实验中，设置 $N=10$，即推荐系统将向用户列出 10 条推荐的推文。将不同的用户兴趣建模策略与上述轻量级的推荐算法一起应用以提供个性化的推文推荐。与先前一些研究[15,18-20,22]中使用的指标相同，前条推文推荐的质量通过以下指标进行衡量。

（1）MRR（Mean Reciprocal Rank），表示用户点赞的推文平均出现在推荐推文列表中的排名。

（2）S@N（Success at rank N），表示用户点赞的推文出现在前 N 个推荐推文中的平均概率。

（3）P@N（Precision at rank N），表示前 N 条推荐推文中被检索到是用户点赞推文的平均概率。

（4）R@N（Recall at rank N），表示用户点赞推文在前 N 条推荐推文中被检索到的平均概率。

4.3　实验结果分析

首先，对用户关注的不同类型账号的数量分布进行分析。如图 2 所示，关注的账号数量低于 100 的用户仅有 86 位，约占 10.4%。用户平均关注的账号数量为 531，其中认证账号数量和名人类账号数量分别为 97 和 51，分别约占用户关注账号的 18%和 10%。

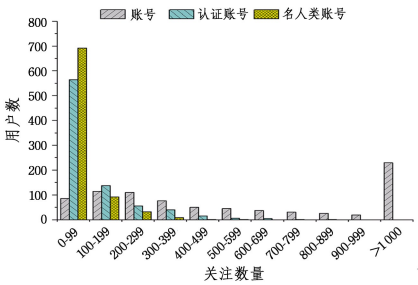

图 2　用户关注账号数量分布

Fig. 2　Distribution of number of user's following accounts

（1）活跃用户兴趣画像构建的效果

为了评估本文提出的兴趣画像模型对于 Twitter 活跃用户的兴趣画像构建效果，比较了两种基于推文的 baseline 方法。与本文相同，这两种方法都使用实体相关的维基百科类别生成用户兴趣画像。另外，这两个方法都采用了一种维基百科类别权重更新策略，公式如下：

$$CategoryDiscount=\frac{1}{\alpha}\times\frac{1}{\lg(SP)}\times\frac{1}{\lg(SC)} \tag{13}$$

其中，SP 为属于该类别的页面集合，SC 为属于该类别的子类别集合。第一种对比方法[16]记作 um(CF)，该方法使用 CF (Category Frequence)作为加权方案，其同样也是文献[17]中对比的 baseline 方法。第二种对比法[17]记作 um(CF-IDF)，该方法使用 CF-IDF(Category Frequence - Inverse Document Frequency)作为加权方案。

图 3 给出了根据不同评估指标，使用不同的活跃用户兴趣建模策略的推荐结果。相比于 um(CF-IDF)，本文方法 PPR(tweets)在各项评估指标上都有提升，其中 MRR 提升了 7%，S@10 提升了 5%，P@10 和 R@10 提升了 4%。

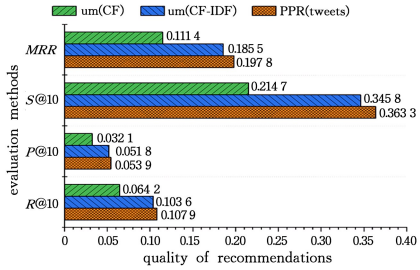

图 3　活跃用户兴趣画像模型评估对比

Fig. 3　Evaluation of active user interest profile model

（2）非活跃用户兴趣画像构建的效果

为了评估本文提出的兴趣画像模型对于 Twitter 非活跃用户的兴趣画像构建效果，比较了另外两种基于关注账号信息的 baseline 方法。第一种对比方法[13]记作 SA(followees_name)，其利用关注用户的名称进行用户兴趣建模。第二种对比方法[15]记作 SA(followees_bio)，其利用关注用户的描述进行用户兴趣建模。这两种方法都使用实体相关的维基百科类别来对用户兴趣进行丰富，并通过传播激活的方式获得由维基百科类别表示的用户兴趣画像。

图 4 给出了根据不同评估指标，使用不同的非活跃用户兴趣建模策略的推荐结果。相比于 SA(followees_bio)，本文方法 PPR(followees)在各项评估指标上都有着明显的提升，其中 MRR 提升了 20%，S@10，P@10 和 R@10 都提升了 9%。

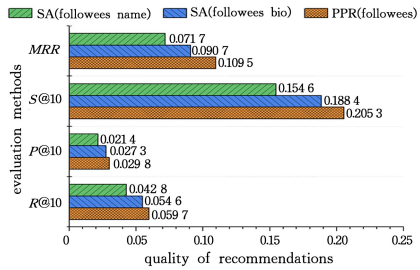

图 4　非活跃用户兴趣画像模型评估对比

Fig. 4　Evaluation of passive user interest profile model

总体而言，在 Twitter 用户兴趣画像构建方法的实验对比中，本文提出的方法在所有评估指标方面的表现都是最好的。其主要原因如下：1）所有的对比方法中都没有考虑到用

户对于同一实体所属不同维基百科类别的兴趣偏好，因此与本文方法相比，其生成的兴趣权重之间的差异不够准确；2)在用户兴趣的拓展方面，um(CF)和 um(CF-IDF)都只使用了实体直接所属的维基百科类别，没有利用维基百科类别图进一步拓展用户兴趣。而 SA(followees_name)和 SA(followees_bio)虽然通过传播激活的方法对用户兴趣进行了拓展，但是在从子类别向父类别的传播过程中使得越靠近顶层的类别的权重越大，从而导致用户之间的兴趣差异较小，因此推文推荐的效果较差。

结束语 本文关注于 Twitter 用户的兴趣画像构建，提出了一种基于维基百科类别图的用户兴趣建模方法。该方法通过使用个性化 PageRank 算法在维基百科类别图上随机游走来进行用户兴趣的拓展，以生成维基百科类别表示的用户兴趣画像。该方法通过使用推文和关注账号信息(名称及个人描述)这两种不同的文本内容，来分别实现 Twitter 活跃用户和非活跃用户的兴趣画像构建。为了评价该方法构建的用户兴趣画像质量，对其在推文推荐系统中的推荐效果进行了评估。实验表明，在 Twitter 活跃用户和非活跃用户的兴趣画像构建方面，本文提出的方法在所有评估指标方面的表现都有着显著提升，能够更加有效地挖掘用户的兴趣。

参 考 文 献

[1] ZHOU X,XU Y,LI Y,et al. The state-of-the-art in personalized recommender systems for social networking[J]. Artificial Intelligence Review,2012,37(2):119-132.

[2] QIU Y F,WANG L Y,SHAO L S,et al. User in-terest modeling based on Weibo short text[J]. Computer Engineering,2014,40(2):275-279. (in Chinese)
邱云飞，王琳颖，邵良杉，等. 基于微博短文本的用户兴趣建模方法[J]. 计算机工程，2014，40(2)：275-279.

[3] WENG J,LIM E P,JIANG J,et al. TwitterRank:finding topic-sensitive influential twitterers[C]// Proceedings of the Third ACM International Conference on Web Search and Data Mining. New York:ACM,2010:261-270.

[4] STEYVERS M,SMYTH P,ROSEN-ZVI M,et al. Probabilistic author-topic models for information discovery[C]// Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York:ACM,2004:306-315.

[5] ZHAO W X,JIANG J,WENG J,et al. Comparing twitter and traditional media using topic models[C]// European Conference on Information Retrieval. Berlin Heidelberg: Springer, 2011:338-349.

[6] CHEN J,NAIRN R,NELSON L,et al. Short and tweet:experiments on recommending content from information streams[C]// Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. New York:ACM,2010:1185-1194.

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

[8] LU C,LAM W,ZHANG Y. Twitter user modeling and tweets recommendation based on wikipedia concept graph[C]// Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence. 2012.

[9] MICHELSON M,MACSKASSY S A. Discovering users' topics of interest on twitter:a first look[C]// Proceedings of the Fourth Workshop on Analytics for Noisy Unstructured Text Data. New York:ACM,2010:73-80.

[10] SIEHNDEL P,KAWASE R. TwikiMe!:user profiles that make sense[C]// Proceedings of the 2012th International Conference on Posters & Demonstrations Track-Volume 914. CEUR-WS.org,2012:61-64.

[11] KAPANIPATHI P,JAIN P,VENKATARAMANI C,et al. User interests identification on twitter using a hierarchical know-ledge base[C]// European Semantic Web Conference. Springer,Cham,2014:99-113.

[12] LIM K H,DATTA A. Interest classification of Twitter users using Wikipedia[C]// Proceedings of the 9th International Symposium on Open Collaboration. New York:ACM,2013:22.

[13] BESEL C,SCHLÖTTERER J,GRANITZER M. Inferring semantic interest profiles from Twitter followees:does Twitter know better than your friends? [C]// Proceedings of the 31st Annual ACM Symposium on Applied Computing. New York:ACM,2016:1152-1157.

[14] FARALLI S,STILO G,VELARDI P. Recommendation of microblog users based on hierarchical interest profiles[J]. Social Network Analysis and Mining,2015,5(1):25.

[15] PIAO G,BRESLIN J G. Inferring User Interests for Passive Users on Twitter by Leveraging Followee Biographies[C]// European Conference on Information Retrieval. Springer, Cham, 2017:122-133.

[16] KENTER T,RIJKE M D. Short Text Similarity with Word Embeddings[C]// ACM International on Conference on Information and Knowledge Management. New York: ACM, 2015: 1411-1420.

[17] GOLDBERG Y,LEVY O. word2vec Explained:deriving Mikolov et al.'s negative-sampling word-embedding method[J]. arXiv:1402.37232014.

[18] PIAO G,BRESLIN J G. Analyzing Aggregated Semantics-enabled User Modeling on Google+ and Twitter for Personalized Link Recommendations [C]// Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization. New York:ACM, 2016:105-109.

[19] PIAO G,BRESLIN J G. Exploring Dynamics and Semantics of User Interests for User Modeling on Twitter for Link Recommendations[C]// International Conference on Semantic Systems. New York:ACM,2016:81-88.

[20] ZARRINKALAM F,KAHANI M,BAGHERI E. Mining user interests over active topics on social networks[J]. Information Processing & Management,2018,54(2):339-357.

[21] FOGARAS D,RÁCZ B,CSALOGÁNY K,et al. Towards scaling fully personalized pagerank:Algorithms,lower bounds,and experiments[J]. Internet Mathematics,2005,2(3):333-358.

[22] ABEL F,HAUFF C,HOUBEN G J,et al. Leveraging user modeling on the social web with linked data[C]// International Conference on Web Engineering. Springer-Verlag,2012:378-385.