文章编号：1003-0077(2018)07-0109-07

面向领域的高质量微博用户发现

叶永君[1,2]，李 鹏[1,2]，周美林[1,2]，万仪方[1,2]，王 斌[1,2]

（1. 中国科学院 信息工程研究所，北京 100093；
2. 中国科学院大学，北京 100049）

摘 要：在微博系统中，寻找高质量微博用户进行关注是获取高质量信息的前提。该文研究高质量微博用户发现问题，即给定领域词查询，系统根据用户质量返回相关用户排序列表。将该问题分解成两个子问题：一是领域相关用户的检索问题，二是微博用户排序问题。针对用户检索问题，提出了基于用户标签的用户表示方法以及基于维基百科的查询—用户相似度匹配方法，该方法作为 ESA(explicit semantic analysis)的一个扩展应用，结果具有良好的可解释性，实验表明基于维基百科的效果要优于基于其他资源的检索效果。针对用户排序问题，提出了基于图的迭代排序方法 UBRank，在计算用户质量时同时考虑用户发布消息的数量和消息的权威度，并且只选择含 URL 的消息来构建图，实验验证了该方法的高效性和优越性。

关键词：用户质量测量；用户行为模型；图排序算法

中图分类号：TP391　　**文献标识码**：A

Domain Specific High-quality Microblogging User Detection

YE Yongjun[1,2], LI Peng[1,2], ZHOU Meilin[1,2], WAN Yifang[1,2], WANG Bin [1,2]

(1. Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China;
2. University of Chinese Academy of Sciences, Beijing 100049, China)

Abstract: In nowadays microblogging systems such as Twitter or Weibo, searching high quality users to follow is essential for acquiring information. This paper is focuses on the task of high quality user identification, i.e., given a domain query, return a user list according to user quality. We divide the task into two sub-problems: the user search and the user ranking. As for the user search, we represent users according to their tags and propose a similarity-based retrieval approach using the Chinese Wikipedia, which is essentially an extension of the current ESA(explicit semantic analysis) method. As for the user ranking, we propose a graph-based ranking method called UBRank, which considers both the quantity and the quality of the published posts to measure the user importance. Experiments indicate that using Chinese Wikipedia is better than other resources such as HowNet, and validate the efficiency and superiority of the ranking method.

Key words: user quality measure; user behavior model; graph based ranking

0 引言

随着信息化进程不断加快，越来越多的普通用户从信息的阅读者演变成了信息的创造者与信息的传播者。其中，微博平台已经成了一个产生热点事件和观察社会言论的重要场所。据估计，Twitter 平台每天有高达五亿条微博消息被人们所发布。这些微博消息主题丰富，既包含一些普通对话，也包含特定领域相关的有价值信息。根据微博系统的功能规则，人们(follower)必须关注其他用户(followee)才能获取信息，这些被关注用户(followee)发布的信息质量完全决定了关注者(follower)所获取信息的价值。考虑到用户价值往往集中在特定领域，选择领域相关的高质量用户[1]进行关注，对于微博使用者进行信息获取具有重要价值：一方面可以获取

收稿日期：2017-08-21　**定稿日期**：2017-11-16

基金项目：国家自然科学基金(61402466)；国家高技术研究发展计划(863)(2015AA016005)

更全面的信息(相关信息),另一方面也可以减轻信息(不相关信息)过载问题。

本文将高质量微博用户发现问题拆解成两个子任务:领域相关用户的检索任务以及用户质量排序任务。领域相关用户检索任务是给定领域,从海量微博用户中找到与该领域相关的用户;用户排序任务是指给定用户集合,根据用户质量对用户进行排序。

在具体方法上,对于领域相关用户检索任务,我们尝试将领域词与微博用户的匹配转化为领域词和用户标签的匹配。其中,为了解决词项失配问题,我们使用了基于维基百科的语义相似度计算方案。该方法首先将领域词、标签词表示为维基百科的词条向量,基于词条向量来计算匹配度。该方法作为ESA(explicit semantic analysis)的一个扩展应用,相比 Word2Vec 或者 LSA 等对应的隐语义,对最后得到的结果有着良好的可解释性。对于用户质量排序任务,我们认为用户质量由用户所发消息质量所决定。进一步的统计分析发现:含 URL 的消息质量更高、对用户表征作用更强,且更容易被转发,应该重点考虑。为此,在计算时我们只考虑含 URL 的消息,构造了基于用户发布关系以及用户转发关系的联合图,通过图迭代得到用户质量以及消息质量得分,基于得分完成用户排序。实验结果表明:该排序方法得到的用户排序结果与基于人工标注得到的用户排序结果具有很高的一致性。

本文后续内容组织如下:第一节介绍相关工作;第二节介绍本文工作;第三节给出实验和结论;第四节对全文进行总结。

1 相关工作

自微博诞生以来,度量用户的重要性一直是一个主要研究问题。相关工作可以分为领域无关用户重要度研究[2-7]和领域相关用户重要度研究[1,8-9]。大部分的研究工作将用户的重要性定义为用户的权威度:即所发信息更容易被转发传播,用户更具有影响力。然而这些工作忽略了用户的信息量,即用户发布的消息数量。实际上,用户发布的高质量消息越多,用户被关注的重要性才越大。目前,考虑用户消息数量的工作只有 Yamaguchi[10]。Yamaguchi 等人的用户测量模型使用用户所有的推文消息来构造 User—Twitter 图。在他们的模型中,用户的消息数量将直接影响用户的测量得分,即在一定程度上,用户所发的微博数量越多,该用户的测量得分会越高。本文与 Yamaguchi 的出发点类似,但存在两方面的不同:本文没有利用用户的全量消息,只利用"含 URL 的消息"来构建图,减少图上的节点数;本文将多种关系进行合并,减少了图的连边。这些改进可以显著加速图的迭代过程。

在计算用户重要度时,相关工作利用的信息包括:用户的关注关系[1,6,8,10-11]、发布行为[10]、转发行为[4,10-11],以及消息内容信息[1,4,11]。具体地,Weng[1]等人提出了 TwitterRank 模型。该算法利用用户之间的关注关系构建有向图,并在该关系图上运行类 PageRank 的算法。Meeyoung 等人的模型利用信息相对较多,专注于三种行为数据:①关注,②转发,③提及 @,并分别分析这三种行为所带来的影响。类似地,Yamaguchi 等人[10]的模型将用户的关注关系、发布行为和转发行为整合到同一个图中;考虑到不同行为的内在意义不同,为不同类型的边设置不同的权重。Gupta 等人[12]的模型中也用到了关注关系,并认为用户之间的关注代表"用户对用户推荐的信任"。上述研究在用户测量时都考虑到了用户之间的关注关系。

2 本文工作

本文方法的整体框架如图 1 所示,输入为用户给定的领域词,输出为与领域相关的高质量微博用户。

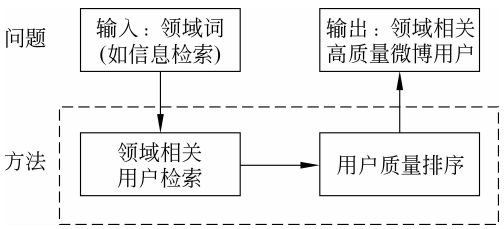

图 1 整体框架

2.1 领域相关用户的检索

我们使用用户标签来表示用户,相关研究表明用户标签对于用户兴趣有很好的指示作用,如 Ghosh S[13]使用 TwitterList 来获取用户兴趣。相应地,在新浪微博平台上,每个用户也会和一个或者一组标签相对应,这里的标签是用户自主标注的,在一定程度上可以反映用户所在领域的信息。以微博用户"爱可可—爱生活"为例(图 2),从标签便能直观地反映用户的领域。

然而利用领域词与用户标签直接匹配会存在

“词项失配”问题，为了提升检索效果，我们借鉴ESA方法[14]的思想，将标签和领域词映射到由维基百科词条构成的高维概念空间中，通过词条向量匹配得到用户与领域的相关度。该方法可以对文本的隐含语义显式表示，便于直观理解向量化后的含义，得到的匹配结果也更容易解释。

图2　微博用户标签示例

2.1.1　外部资源的获取和数据预处理

维基百科页面分为页面网和类别网，本文的研究工作只涉及到页面网。我们下载了最新的Wiki-Dump中文资源数据，对文本进行繁简转换，统一转为中文简体，该WikiDump数据集可以看作是一系列维基百科词条页面的集合。

利用上述方法，我们获得了两组数据。一是基于2015年10月13号的中文维基百科镜像，原始大小为1.2GB，数据处理后得到866 180篇词条文档；二是基于2017年01月02号的中文维基百科镜像，原始大小为1.4GB，数据处理后得到了1 260 760篇词条文档。这些词条涵盖了各方面领域信息。

2.1.2　基于显式语义分析(ESA)的用户检索

基于ESA的用户检索主要分为两步：一是使用ESA方法将领域词和标签词表示为由维基百科概念组成的加权向量(后文称为解释向量)；二是基于解释向量计算领域词与标签词的余弦相似度，取相似度最高的用户作为领域相关用户。

对于ESA方法，具体地，令 $T=\{\varphi_i\}$ 表示输入文本，其对应的TF-IDF向量记作$\langle \boldsymbol{\omega}_i \rangle$，其中 ω_i 是单词 φ_i 的权重。令$\langle k_j \rangle$ 是词 ω_i 对应的维基百科词条，其中 k_j 为输入 φ_i 与维基百科词条 C_j 的关联度，$\{C_j \in C_1,\cdots,C_N\}$(其中 N 表示资源中维基百科词条的总数)。这样的话，对应文本 T 的语义解释向量 $\boldsymbol{V}$ 是一个长度为 N 的向量(对应 N 个词条)，其中每个维基百科词条 C_j 的权重被定义为 $\sum_{\varphi_i} \omega_i \times k_j$。向量 $\boldsymbol{V}$ 的每一维反映了对应的维基百科词条 C_j 与给定文本 T 之间的相关性，如果词条 C_j 与原始文本关联较大，那么对应的特征权重也越大。

表1、表2给出了使用ESA得到的解释向量。以“机器学习”为例，从中我们可以看到解释向量能够对原始词条进行扩展，引入相关特征：部分是与输入词相同或者相近的词条特征，这些特征与输入词存在横向关系，如词条“人工智能”等；二是输入词的上位词或者输入词的下位词，这些特征与输入词存在纵向关系，如“特征缩放”等。显然，通过ESA的转换扩展，可以在一定程度上解决“词项失配”问题。

表1　基于2015年中文维基百科得到的解释向量示例(部分)

序号	机器学习	信息检索	数学	张艺谋
1	机器学习：2.639	文本信息检索：3.402	数学：5.056	张艺谋：3.761
2	深度学习：2.197	信息检索：1.946	新数学：4.094	新画面：2.708
3	人工智能：1.946	词干提取：1.609	四色定理：4.043	归来(电影)：2.639
4	计算机视觉：1.792	停用词：1.386	离散数学：4.025	章子怡：2.398
5	Apache Spark：1.609	搜索引擎：1.386	物理学史：4.007	巩俐：2.398
6	特征缩放：1.386	文本挖掘：1.099	集合论：3.850	李曼：1.946
7	过适：1.098	余弦相似性：1.099	微积分学：3.688	长城(电影)：1.946
8	决策树：1.098	个性化检索：1.099	哥德巴赫猜想：3.688	英雄(电影)：1.792
9	VC理论：1.098	维数灾难：0.693	公理：3.663	周冬雨：1.792
10	强化学习：1.098	相关反馈：0.693	群：3.433	千里走单骑：1.791

表2　基于2017年中文维基百科得到的解释向量示例(部分)

序号	机器学习	信息检索	数学	张艺谋
1	机器学习：2.639	文本信息检索：3.402	数学：5.552	张艺谋：3.761
2	数据挖掘：2.342	信息检索：1.792	新数学：4.094	新画面：2.708
3	深度学习：2.197	停用词：1.609	四色定理：4.043	归来(电影)：2.639
4	人工智能：1.946	词干提取：1.386	物理学史：3.937	章子怡：2.398
5	特征缩放：1.872	搜索引擎：1.386	离散数学：3.937	巩俐：2.398
6	支持向量机：1.716	文本挖掘：1.379	集合论：3.838	长城(电影)：2.076
7	计算机视觉：1.386	SimRank：1.142	微积分学：3.788	英雄(电影)：1.946

续表

序号	机器学习	信息检索	数学	张艺谋
8	过适：1.098	余弦相似性：1.099	传统数学：3.688	李曼：1.942
9	决策树：1.098	个性化检索：1.099	公理：3.663	周冬雨：1.792
10	VC理论：1.098	维数灾难：0.693	群：3.433	千里走单骑：1.791

表1向量化所用资源为2015年10月13号对应的866 180条词条文档数据。为了说明随着维基百科资源的扩大，词汇量的增加可以提高向量化的效果，本文再采用2017年01月02号对应1 260 760条词条的文档数据进行同样的向量化处理，得到的结果如表2所示。对比表1和表2可知，随着资源的更新和扩充，同一输入文本对应向量会发生些许变化，比如机器学习对应向量中词项top10中新增了“数据挖掘”词条，可见近些年，数据挖掘和机器学习两者交叉的越来越多；特别对于输入词“张艺谋”向量词项“长城(英雄)”的权重有所提升，这和“长城”电影刚上映这一热点事件相对应。从新旧向量对比可知，随着内容的更新和增加，新向量能在一定程度上反映一些热点事件、新事件。后文将通过实验来观察新旧资源对实验结果的影响。

2.2 用户质量排序

该任务的输入是2.1节返回的领域相关用户集合，输出是用户质量排序结果。用户排序可以通过计算用户质量得分来解决。现有工作大部分利用用户的关注关系以及消息转发关系，通过构造相应关系图进行图排序得到用户排序结果。实际上，现有方法存在以下两方面的问题：大部分现有方法可以识别高权威(authority)用户，但不能识别高信息量(hub)用户，而对于信息获取需求来讲，用户发布的消息数量与消息质量在衡量用户重要度上是同等重要的；②并不是用户的所有消息都是高质量的，在计算用户重要度时，简单考虑用户发布的所有消息会引入极大的计算量。

基于上述两个问题，我们探索只使用含URL的用户消息以及消息转发关系来对用户质量进行排序。具体地，我们首先验证了含URL的消息相比不含URL的消息，其消息质量更高，更容易被转发，只使用含URL消息计算用户质量可以显著减少计算量；接着我们提出了一种基于图的用户排序方法UBRank(URL biased User Rank)，图中只包含含URL的消息节点，利用消息发布以及消息转发关系来迭代计算用户以及消息的重要度。

2.2.1 含URL消息的统计分析

为了考察消息质量与是否包含URL的关系，我们从3.1的数据集中随机抽样了60个用户(根据用户发布的消息数量切分为六个区间，切分点为100、500、1 000、2 000、5 000，每个区间抽样10个用户)，对每个用户，随机抽取20条含URL的消息以及20条不含URL的消息进行人工标注。消息质量使用三个标注级别：0表示与用户标签不相关，1表示相关，2表示相关且有趣。

令Θ_{URL}表示含URL的消息质量平均得分，$\Theta_{\overline{\text{URL}}}$表示不含URL的消息质量平均得分，基于人工标记结果，在60个用户上进行宏平均，得到$\Theta_{\text{URL}}=0.93$，$\Theta_{\overline{\text{URL}}}=0.29$；对应的双样本t检验(paired t-test)说明含URL的消息质量要显著高于不含URL的消息质量。这个结论也与观察一致，不含URL的消息往往表达情感以及个人生活，与用户兴趣的相关度较低；含URL的消息往往讨论相关的新闻、观点，与用户的兴趣更加一致。

进一步分析发现：对于含URL的消息，11.6%的消息被转发；对于不含URL的消息，4%的消息被转发。这表明含URL的消息包含更多用户交互行为，更容易计算其质量。上述分析有效地说明了只利用含URL的消息来度量用户重要度的合理性。

2.2.2 UBRank图结构

在计算用户质量得分时，我们使用如下假设：①用户发布消息被其他高质量用户转发越多，那么用户质量也越高；②用户发布高质量消息越多，那么用户质量也越高；③消息被高质量用户转发越多，那么消息质量也越高。

基于上述假设，我们将“用户—用户”转发图以及“用户—URL消息”发布图合并为一个统一的图，基于该图来计算用户质量，图中的消息节点为包含URL的消息，而非所有的消息。

具体地，UBRank的图结构如图3所示。

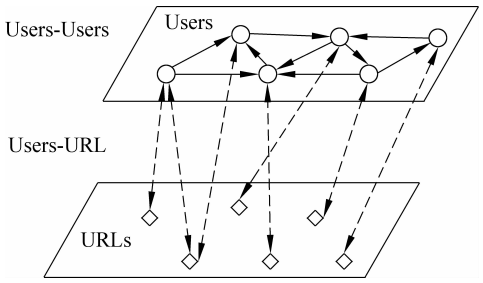

图3 UBRank的双层图结构

从图 3 中可以看到：①用户节点入度来自用户和 URL；②URL 节点入度来自用户。这与我们前边假设一致。

关于图中边的含义：①用户与用户之间的有向边，代表转发关系；②用户与 URL 之间的双向边，代表发布(含转发)关系。

2.2.3　UBRank 迭代算法

假设用户质量得分为 $\boldsymbol{v}=[\upsilon(s_i)]_{m\times 1}$，消息质量得分为 $\boldsymbol{v}=[\nu(t_j)]_{n\times 1}$。UBRank 的迭代公式如式(1)、式(2)所示。

$$\upsilon(s_i)=\alpha\sum_{j=1}^{m}U_{ji}\times\upsilon(s_j)+\beta\sum_{j=1}^{n}V_{ji}\times\nu(t_j) \quad (1)$$

$$\nu(t_j)=\sum_{i=1}^{m}V_{ij}\times\upsilon(s_i) \quad (2)$$

其中矩阵 $\boldsymbol{U}$ 对应用户—用户图，矩阵 $\boldsymbol{V}$ 对应用户—URL 图。用户的质量得分由其相邻用户以及发布的消息质量得分决定，含 URL 消息的质量仅仅由其相邻的微博用户决定。相应的矩阵形式可表示为式(3)、式(4)。

$$\boldsymbol{\upsilon}=\boldsymbol{\alpha U}^{\mathrm{T}}\boldsymbol{\upsilon}+\boldsymbol{\beta V}^{\mathrm{T}}\boldsymbol{\nu} \quad (3)$$

$$\boldsymbol{\nu}=\boldsymbol{V}^{\mathrm{T}}\boldsymbol{\upsilon} \quad (4)$$

其中 α 和 β 分别表示来自同质节点和异质节点(类似 Hits 算法中的 Hub 和 Authority 节点)对最终质量得分的相对贡献程度，$\alpha+\beta=1$。为了保证迭代收敛，每轮迭代结束时 $\boldsymbol{\upsilon}$ 和 $\boldsymbol{\nu}$ 都要进行归一化。

3　实验和结论

3.1　实验准备

为了验证领域相关用户检索方法的有效性，首先需要一组微博用户集合以及对应的用户标签。本文通过获取种子用户的两层关注数据，采集到了 21 042 个不同用户，这些用户属于各个领域。通过进一步用户分析，我们发现其中 16 571(占总体用户的 78.75%)个用户拥有标签数据，本文使用该 16 571 个用户及其标签的集合作为本文的实验室数据集。

3.2　领域相关用户检索—对比实验设置

为了验证本文提出的基于维基百科的显式向量表示法的有效性，我们选择领域查询“机器学习”和“信息检索”，比较不同方法得到的领域用户集合的相关性。具体地，我们实现对比了以下几种用户检索方法。

(1) 基于维基百科 ESA 的相似计算法：如前面方法分析所述，利用维基百科页面网的词条文档对领域词和标签进行向量化，这里本文有 2015-10 和 2017-01 两份资源，分别记作维基 15 和维基 17。其中利用倒排索引获得对应词条的权重后，为了去除噪音和不重要的关联关系，按照词条权重排序，只保留排名最高的前 80%的词条。

(2) 基于知网的语义相似度计算法：利用知网中的义原对词语进行解释，并基于义原进行相似度计算，该方法简称为“知网”。

(3) 基于 Word2Vec＋中文维基百科资源的语义相似度计算方法：利用 Word2Vec 框架训练中文维基百科资源，此处直接用最新的 2017-01 对应的维基百科资源。训练方式选择的 CBOW，该方法简称为“Word2Vec”。

3.3　领域相关用户检索-结果与评价

正如前文所述，该部分问题是一个典型的信息检索问题，已知领域词，得到匹配的用户集合。考虑到人工标注的耗时和高成本，本文仅仅使用正确率(Precision)作为评价指标。具体来说，统计各个实验结果的 P_5、P_{50}，P_{100} 和 P_{200}。实际操作层面，本文至多只需要标注各个实验的 top200 即可。经过 pooling 后，针对领域词“机器学习”和“信息检索”，实际本文分别只得到了 429 个和 447 个不同的用户，只需要人工标注这些用户即可。评价结果如表 3、表 4 所示。

表 3　领域词“机器学习”检索效果/%

	P@5	P@50	P@100	P@200
知网	100	82.0	77.0	61.0
Word2Vec	100	86.0	82.0	66.0
维基 15	100	88.0	84.0	70.0
维基 17	**100**	**92.0**	**87.0**	**74.0**

表 4　领域词“信息检索”检索效果/%

	P@5	P@50	P@100	P@200
知网	100	85.0	80.0	63.0
Word2Vec	100	88.0	85.0	67.0
维基 15	100	92.0	86.0	74.0
维基 17	**100**	**93.0**	**89.0**	**77.0**

从表 3 和表 4 可以看到，维基 15 和维基 17 要优于其他方法，说明基于维基百科 ESA 的相似度计算方法的有效性。再对比这二者可知，2017 年的数据集效果明显优于 2015 年的数据集，说明随着资源

规模的扩大,检索效果会有进一步提升。

3.4 用户质量排序—对比实验设置

为了验证本文所提的 UBRank 排序方法的有效性,本文实现了以下五种用户排序方法。

(1) UBRank:如前面算法分析所述,UBRank 只关注含 URL 的消息,并基于用户—用户转发图和用户—URL 发布图进行图算法构建。通过训练所知,参数 α 和 β 都设置为 0.5。

(2) RTRankU:此方法仅仅基于"含 URL 消息"的转发消息构建用户—用户转发图,此时忽略用户—URL 发布图。本文将在此用户—用户转发图上运行 PageRank 算法。

(3) RTRankA:此方法基于所有消息的转发关系构建用户—用户转发图,并依旧忽略用户—URL 发布图。本文也在此完全的用户—用户转发图上运行 PageRank 算法。

(4) TuRank:TuRank 算法考虑所有的行为数据:关注行为、发布行为和转发行为。图中的节点表示用户,用户—消息之间的发布行为和用户之间的转发行为都会映射到用户之间的边上。并按照文献[10]中工作对"不同关系(粉丝关系和转发关系)边"对应的权重进行区分设置。此方法是目前相关工作中表现最好的模型,是本文模型的重点参照对象。

(5) TwitterRank:这个模型是文献[1]中算法的简化版本。本文跳过了从消息中计算用户主题因子的过程,因为从一开始,本文挑选出的用户集合已经限定在某个特定的主题中。该方法仅仅基于用户之间的关注关系构建用户—用户关注图。

3.5 用户质量排序-结果与评价

为了度量算法效果,我们通过人工标记获得用户质量标准序,通过比对算法得到的序与标准序的差异来评估算法效果。序的度量使用 Kendall's τ [15]作为评估指标。τ 值越大意味着算法得到的序越接近人工判断。

在进行人工标注时,我们先对用户消息进行标注,用户质量得分等于用户消息得分的累加和。考虑到用户发布消息规模很大,为了降低标注量,我们对用户消息进行分层抽样,只对抽样结果进行标注,基于样本标注得分来估算用户所有消息得分。具体地,对每个用户,根据消息<是否含 URL、转发量>进行分组。"是否含 URL"对应二类:包含 URL、不含 URL,转发量分为三个区间:[0,1),[1,5),[5,+∞),所以每个用户的消息被划分为六组。我们在每组抽样五条消息进行标注,每个用户平均有 30 条消息被标注。我们将消息质量划分为三个等级:0 表示领域不相关,1 表示相关,2 表示相关且有趣。各算法得到的用户排序性能如表 5 所示。

表 5 Top 10 用户实验结果对比

模型	Kendall's τ
UBRank	**0.946**
RTRankU	0.864
RTRankA	0.812
TuRank	0.842
TwitterRank	0.794

从表 5 可以发现,本文所提的 UBRank 要优于其他对比实验。RTRankU 的效果要好于 RTRankA,说明了"仅使用带 URL 消息"相比使用"所有消息"在计算用户质量排序上的优越性。具体地,在图构建上,RTRankU 只使用带 URL 的消息,而 RTRankA 使用所有消息,其中包括了那些不带 URL 的消息。除了这点不同外,其他的过程对于两种方法是完全相同的。这也与前面的统计分析相一致:含 URL 的消息比不含 URL 的消息质量更高;反过来说,不含 URL 的消息由于转发量有限且话题无关,对于用户质量测量可能引入噪音。

从表 5 中我们还可以发现 RTRankA 的实验效果与 TwitterRank 的效果相当;这表明转发关系与关注关系在计算用户重要度上效果相当。TuRank 的实验效果优于 RTRankA、TwitterRank,这一结果表明,通过组合关注信息和转发信息可以提升实验效果。RTRankU 同时优于 RTRankA、TwitterRank 表明:相比利用"所有消息的转发"和"用户之间的关注"信息,利用"含 URL 消息的转发"信息计算用户质量更为有效。此外,对数据集中所有用户(21 042 用户)的所有消息进行统计,我们发现含 URL 的消息量只占总体消息量的 20%,利用含 URL 消息计算用户质量可以极大地减少计算规模。

4 总结与展望

本文研究面向领域的高质量微博用户发现问题,并将该问题分解为两个子问题:领域相关用户的检索以及用户质量排序。对于领域相关用户检索,我们使用用户标签来表示用户,通过计算领域词

与用户标签的匹配度，取排名最高的用户作为领域相关用户，领域词与用户标签匹配使用基于中文维基百科的显式向量（ESA）的语义相似度计算方法，实验验证了 ESA 方法在检索领域相关用户方面的有效性和优越性，并通过 2015 年和 2017 年新旧资源对比，说明随着资源的更新，匹配精度会得到进一步提升。对于用户质量排序，我们提出了基于图的迭代排序方法 UBRank，在计算用户质量时同时考虑用户发布消息的数量和消息的权威度，并且只选择含 URL 的消息来构建图，实验表明仅使用含 URL 的消息相比使用全部消息得到的用户质量排序效果更好，并且引入的计算规模更小。

未来的工作包括：通过引入更多中文资源来提升语义相似度的匹配效果、对 URL 做进一步过滤、考虑引入时间因素对用户质量进行评价等。

参考文献

[1] Jianshu Weng, Ee-Peng Lim, Jing Jiang, et al. Twitterrank: finding topic-sensitive influential twitterers [C]//Proceedings of the third ACM International Conference on Web Search and Data Mining, 2010: 261-270.

[2] Eytan Bakshy, Jake M Hofman, Winter A Mason, et al. Everyone's an influencer: Quantifying influence on twitter[C]//Proceedings of the fourth ACM International Conference on Web Search and Data Mining, 2011: 65-74.

[3] Changhyun Lee, Haewoon Kwak, Hosung Park, et al. Finding influentials based on the temporal order of information adoption in Twitter[C]//Proceedings of the 19th International Conference on World Wide Web, 2010: 1137-1138.

[4] Daniel M Romero, Wojciech Galuba, Sitaram Asur, et al. Influence and passivity in social media[C]//Proceedings of Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, 2011: 18-33.

[5] Meeyoung Cha, Hamed Haddadi, Fabricio Benevenuto, et al. Measuring user influence in twitter: The million follower fallacy[J]. Icwsm, 2010, 10(10-17): 30.

[6] 彭泽环，孙乐，韩先培，等. 基于排序学习的微博用户推荐[J]. 中文信息学报，2013，27(04)：96-102.

[7] 张绍武，尹杰，林鸿飞，等. 基于用户分析的微博用户影响力度量模型[J]. 中文信息学报，2015，29(04)：59-66.

[8] D Tunkelang. A twitter analog to pagerank. The Noisy Channel, 2009[DB/OL]. http://thenoisychannel.com/2009/01/13/a-twitter-analog-to-pagerank.

[9] Shoubin Kong, Ling Feng. A tweet-centric approach for topic-specific author ranking in microblog[C]//Proceedings of International Conference on Advanced Data Mining and Applications, Springer, 2011: 138-151.

[10] Yuto Yamaguchi, Tsubasa Takahashi, Toshiyuki Amagasa, et al. Turank: Twitter user ranking based on user-tweet graph analysis[C]//Proceedings of International Conference on Web Information Systems Engineering, Springer, 2010: 240-253.

[11] Aditya Pal, Scott Counts. Identifying topical authorities in microblogs[C]//Proceedings of the fourth ACM International Conference on Web Search and Data Mining, 2011: 45-54.

[12] Pankaj Gupta, Ashish Goel, Jimmy Lin, et al. Wtf: The who to follow service at Twitter[C]//Proceedings of the 22nd International Conference on World Wide Web, 2013: 505-514.

[13] Saptarshi Ghosh, Naveen Sharma, Fabricio Benevenuto, et al. Cognos: Crowdsourcing search for topic experts in microblogs[C]//Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2012: 575-590.

[14] Evgeniy Gabrilovich, Shaul Markovitch. Computing semantic relatedness using wikipedia-based explicit semantic analysis[C]//Proceedings of the 20th IJCAI, 2007: 1606-1611.

[15] Maurice George Kendall. Rank correlation methods [J]. Journal of the Institute of Actuaries, 1949 (1): 140-141.

叶永君（1993—），硕士研究生，主要研究领域为信息检索。

E-mail：evapandora@sina.cn

周美林（1990—），硕士，助理研究员，主要研究领域为信息检索。

E-mail：zhoumeilin@iie.ac.cn

李鹏（1985—），通信作者，博士，高级工程师，主要研究领域为信息检索、自然语言处理。

E-mail：lipeng@iie.ac.cn