基于领域划分的微博用户影响力分析

刘金龙 吴 斌 陈 震 沈崇玮
(北京邮电大学计算机学院 北京 100876)

摘 要 近年来微博作为一种新兴的社交网络逐渐被广大用户使用。微博信息简短、更新迅速、包含信息量大,给微博用户获取信息带来了诸多不便,因此,利用影响力分析的手段找到具有较大影响力的微博用户具有重大意义。微博内容较传统的媒体信息具有较强的时效性和权威性,同时微博用语也极其不规范,这给微博用户影响力的分析带来了极大的困难。首先对获取的微博用户信息进行领域的划分,采用基于微博内容和用户关注的方式将用户归类到其所属的领域。其中,采用新词发现以及特征扩展的方法来提高划分结果的准确性。然后,对各个领域的用户进行影响力分析,提出3种影响力传播模型,用户最终的影响力大小根据3种模型的结果进行加权计算。最后对实验结果进行分析、比较,证明了计算用户影响力的方法能取得较优的结果。

关键词 新浪微博,领域划分,影响力,文本分析,新词发现
中图法分类号 TP393 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2015.5.008

Research on Influence of Micro Blogging Based on Field Division

LIU Jin-long WU Bin CHEN Zhen SHEN Chong-wei
(College of Computer Science, Beijing University of Posts and Telecommunications, Beijing 100876, China)

Abstract In recent years, as an emerging social network, Microblog is being used by a lot of users. The micro blogging platform contains a large amount of information and the update speed of the information is fast so that it often makes users can not find the information they need. Then, it is important to help users find the information which is sent by people who have a great influence. The content that the micro blogging platform publishes and updates is very fast, and the content is not standardized, so that the timeliness and authoritativeness tend to be more important. The purpose of research on influence of micro blogging users in this paper is to identify the influence of different users in different areas. This paper divided the users into different areas by using two different features which are the content of the micro blogging and the concerned relationship of the users. During this, we used the parallel new word recognition algorithm in the micro blogging content and semantic extensions on some important text feature to improve accuracy of the dividing result. Then we computed the influence of the users in all the classification using three models, and combined the result in different weight. At last, we tested the accuracy of the result and compared it with other ways.

Keywords Sina microblog, Field division, Influence, Text analysis, New word recognition

Web2.0时代,互联网正逐渐取代传统的社交媒体,成为人们与外界进行交互的主要渠道。近年来,微博作为一种随着Web2.0而新兴的社交网络正在不断发展。微博以其独特的应用传播模式,在社会网络中产生了巨大的影响力,截止到2013年,新浪微博注册用户达5亿。微博内容短小精悍,新浪微博限制一条微博在200字以内,简短的语句可以强烈地表达用户的看法;另外,微博传播方式多种多样,微博可以快速、便捷地通过各种客户端传播,利用被动阅读的传播方式,可以更加及时地被用户读取。微博的迅速发展已经使它超越了信息传播与分享平台的作用,逐渐成为一种新的互联网生态系统。

微博中用户数量不断增长,必然会有一部分影响力较大的用户在微博信息的传播中起到导向作用,从而引起广大用户的关注,因此,影响力较大的用户在微博中具有极大的权威性。另外,影响力较大的用户在不同的领域中往往起着不同的作用,对于普通的网民来说,他们更倾向于信任那些在某一领域中的权威微博用户。因此,按照领域的划分,寻找出不同领域中影响力较大的用户,对于微博信息的分析和挖掘有着极大的作用。

诸多研究者对于社交网络中用户影响力的分析已经进行了深入的探索。研究者在对Twitter的分析中发现少量权威用户能够吸引大量的普通用户的关注和评论,而权威用户更

到稿日期:2014-02-18 返修日期:2014-04-14 本文受国家重点基础研究发展计划(973)(2013CB329606),国家自然科学基金项目(71231002)资助。

刘金龙(1991-),男,硕士,主要研究方向为智能信息处理、数据挖掘、云计算,E-mail:1054986434@qq.com;吴 斌(1969-),男,博士,教授,主要研究方向为智能信息处理、数据挖掘、云计算;陈 震(1990-),男,硕士,主要研究方向为智能信息处理、数据挖掘、云计算;沈崇玮(1989-),男,硕士,主要研究方向为智能信息处理、数据挖掘、云计算。

倾向于关注同一种类型的用户[1]。社交网络用户之间的交互行为构成了一张复杂的社会网络，在之后的研究中，研究者们以网络中的人度以及用户之间的评论、转发等交互行为作为网络中的边权，利用 PageRank 网络传播算法分析社交网络中用户的影响力[2]，取得了较好的结果。结合社交网络的独特性质，研究者们对 PageRank 方法进行了诸多的改进，利用主题敏感 PageRank[3] 等方法，将情感分析、文本分析等方法引入用户影响力的计算中，识别出其中的意见领袖。其它的研究方法，如 HITS(Hypertext-Induced Topic Search)等经典的链接分析算法也被不少分析者采用，但 HITS 算法由于在初始种子用户以及影响力传播中的局限性[4]，与 PageRank 的方法相比，效果较差。唐杰在研究中提出了基于主题的影响力传播模型，以此来衡量用户在不同主题下的影响力[5]。

根据微博等社交网络的特点，文献[6]利用协同过滤和文本聚类的方式将微博用户进行分类，发现有着相同兴趣爱好的用户通常关注相同的主题。针对微博内容进行话题检测，识别出微博中的不同主题，从而达到用户划分领域的方法也被采用的目的。文献[7]采用改进的 Tf/Idf 对微博内容进行建模，根据微博内容进行聚类。另外，基于文本分类的方式，研究者也对诸如 KNN(K-Nearest Neighbor)[8]、决策树、SVM(Support Vector Machine)[9] 等方法进行了深入的分析。

本文根据之前的研究基础，利用获取的新浪微博信息，对微博用户进行领域划分，然后识别出不同领域中影响力较大的用户信息。用户的领域划分实质上是指将不同兴趣的用户进行分类。

1 基于微博内容和用户关注的微博用户领域划分

1.1 基于微博内容的用户领域划分

本文研究的核心是对微博用户进行影响力的分析。众所周知，微博用户关注的领域不同，在不同的领域中的影响力也会有很大的差别。微博内容是用户对某一话题信息发表的看法和评论，在某一领域中具有较强影响力的微博用户往往会更加关注其所属领域中的信息动态，因此，对于表征用户领域具有重要意义。利用微博内容对微博用户进行领域划分是本文的一个研究重点。

利用微博内容进行用户领域的划分主要是采用文本聚类的方法对用户的微博内容进行聚类，发现用户的领域特征。本文采用中科院的 ICTCLAS 分词器对微博内容进行分词，由于不同的特征词语对领域划分的贡献不同，对于分词特征采取加权方式。过程中采用卡方检验算法对高维的词语特征进行特征选择及降维处理。微博内容比较简短，且微博中网络新词、命名实体等词较多，对文本聚类的结果影响较大，针对这些不利因素，本文采用多种方法去矫正，以提高结果的准确性。利用 ICTCLAS 可以比较准确地识别微博内容中的命名实体；另外，本文采用一种网络新词发现的算法来识别微博文本中的网络新词，这种方法采用 Accessor Variety(简称 AV 值)[10] 的概念衡量字串成词的可能性。算法分别统计字串左侧和右侧紧邻的不同字符个数，记为 L_{av}、R_{av}，而字串的 AV 值记为 $\min(L_{av}, R_{av})$，字串的 AV 值高于一定的阈值，则认为该字串形成新词。微博内容较短，在领域划分时，这种较短的文本内容很难完全表达领域的特征，利用特征扩展的方式，可以对微博文本中重要的特征词语进行扩展，扩大特征词语的覆盖，提高领域的识别程度。具体利用搜索引擎提供的“相关搜索”来对微博词语进行扩展，相关搜索是搜索引擎为了满足用户搜索需求，根据用户的查询和点击行为进行精准的匹配和推荐而产生的，利用相关搜索可以得到与查询词关系紧密的词语扩展。通过收集搜索引擎中的相关搜索，选择扩展词语中频度较高的词语，对微博特征空间进行扩展，丰富词语特征的领域覆盖信息。采用 $tf * idf$ 词语权重计算方法对特征选择的词语进行权重计算，组成微博内容向量。

最后，利用 K-means 文本聚类的方式对微博文本内容进行领域的划分，计算每个微博用户到各个类别中心的相似度，并进行归一化处理，作为用户在不同领域中的权重。为评估聚类结果的准确性，本文首先抽出微博用户中所属领域明显的粉丝数量超过 10 万的热门用户，从网络中抽取出不同领域中的热门人物名单，结合微博数据信息人工地进行标注，共抽取 500 个热门用户，按照新浪自身的领域划分方法，分属体育、政治、娱乐、教育、汽车、旅游、IT、游戏、生活 9 个领域。根据对人工标注的热门用户划分的结果的准确率判断算法在整个数据集中的结果。

1.2 基于用户关注的用户领域划分

微博用户之间的粉丝关注关系是微博中比较稳定的用户关系，这种稳定的用户关系在表达用户所属领域时具有较强的作用，通常关注某一领域中的微博用户数量较大的用户往往也属于该领域。与协同过滤的方法类似，基于用户关注的领域划分将关注的微博用户当作物品(Item)，相同关注用户较多的微博用户之间通常存在着相似性，属于同一领域；与协同过滤的方法类似，这种方法同样存在冷启动的问题，对于热门的微博用户可以起到很好的效果，而利用普通用户的共现去寻找同一领域的微博用户则需要依靠用户的实际交往圈，因此，针对普通用户和热门用户需要采用不同的划分策略来对微博用户进行领域划分。

方法中，热门以及粉丝数量较多的微博用户对其关注的用户列表采用协同过滤的方法划分其所属领域，普通微博用户则按照包含核心用户的数量判定普通用户领域。本文给出了以下 3 种用户的定义方式。

1. 热门用户：这一类型用户领域比较明显，为大众所熟知的微博大 V。文中关于算法结果的准确率也是以这些用户的划分结果作为依据，通常定义用户粉丝数量超过 5 万的微博用户为热门用户。

2. 核心用户：热门用户一方面数量较少，另一方面会被不同领域的微博用户所关注，不具有区分性。核心用户是指在微博中某一领域具有较强影响力而又不像热门用户那样为大众所熟知的微博用户，这类用户具有极强的领域性，被特定领域的微博用户大量关注，而不同领域的用户则关注较少。定义粉丝数量在 500 以上的微博用户为核心用户，这类用户在各领域中影响力较大。

3. 普通用户：这一类型的用户通常交往圈比较局限，限于自己的亲人、同事、好友等，数量规模比较大。因此利用普通用户的关注度去划分用户领域比较困难，按照其关注列表中核心用户所属领域对这类用户进行划分。

根据这种用户类型的定义，对每一类型的微博用户按照不同的划分方法进行领域的划分，以提高结果的准确性。利用 K-means 算法根据用户关注对微博用户进行领域划分，同

样利用上文中人工标注的热门微博用户检验算法的结果，用户向量采用0、1布尔类型向量根据用户的关注列表建立。计算微博用户到所有领域中心的相似度，将其组成向量，归一化后作为用户的领域权重。

在利用以上两种领域划分的方法对微博用户进行领域划分之后，本文在标注的数据集上进行多组实验，将二者结果加权合并，得到最后的用户领域划分结果，并对结果的准确率进行分析。

2 微博用户影响力的分析

2.1 用户影响力传播模型的建立

微博中用户之间的粉丝关注以及对微博的转发评论等交互行为构成了一张庞大的网络关系。其中微博用户以及微博内容可以视作整个网络中两种不同类型的节点，用户与用户之间的点名、互粉的关系构成了用户之间的连接边。用户发布微博及对微博内容进行转发、评论则构成了用户与微博之间的连接边。由于用户和微博属于两种不同类型的节点，因此整个关系网络便构成了一个复杂的二分图结构，图中包括用户和微博两种类型节点，以及用户与用户、用户与微博两种类型的边结构，根据二分图结构特征，定义其为用户-微博-用户模型图，记为(U-W-U图)。本文采用加权影响力分析算法在用户-微博-用户模型图中计算不同领域中用户的影响力。

本文计算用户影响力的过程分为3个部分。

1. 利用用户-用户加权图计算用户的影响力，这一步主要是构建用户联系网络图，根据用户粉丝的质量和数量来衡量用户影响力的大小，具有高质量粉丝的用户本身必然也具有权威性，影响力较大。用户与用户之间关系的加权采用如下3种策略：

1)假设用户 i 在微博中采用@点名的方式，对用户 j 点名次数为 A_{ij}，而收集到的用户 i 所有点名次数总和为 A，则对用户 i 和 j 边的加权因子记为 $\frac{A_{ij}}{A}$；

2)利用用户领域划分的方法，将计算得到的用户 i 和用户 j 在各个领域中权重的归一化向量作余弦相似度计算，得到用户 i、j 的相似度值为 S_{ij}，则对用户 i 和 j 边的加权因子记为 $\frac{S_{ij}}{\sum S_{in}}$，其中 n 为用户 i 的邻接点集合。

3)根据用户 i 在领域 k 的权重 F_{ik}，用户 i 对用户 j 在领域 k 的影响力 F_{jk}，用户 i 和 j 边的加权因子记为 $T_{F_k}=\frac{1-\log(|F_i(k)-F_j(k)|+1)}{N-\sum_{n\in N(i)}\log(|F_i(k)-F_n(k)|+1)}$，其中 n 为用户 i 的邻接点集合。

以上3种加权因子分别记为 T_1、T_2、T_3，则用户 i 和 j 之间的传播权重记为 $\sqrt[3]{T_1*T_2*T_3}$。

2. 利用用户-微博加权图计算用户的影响力。这一部分通过构建用户、微博之间的转发评论二分图，根据用户发布的微博的质量，促进用户的影响力。用户-微博网络图中，微博用户与用户发布的微博各自成为两种不同类型的网络节点，两种不同的由用户节点 i 发布的微博节点 v 之间生成有向边 $v\rightarrow i$，而用户节点 j 评论或者转发微博节点 v 会产生有向边 $j\rightarrow v$，这样微博用户、微博内容根据边的方向和权重不断传播权重，直到形成稳定的网络结构。模型中微博-用户边 $v\rightarrow i$ 的权重均记为1，这样微博用户 i 发布的微博产生的影响力可以快速地反馈给 i。用户-微博边 $j\rightarrow i$ 的权重按照下述方法加权：

假设用户 j 所有的转发评论微博次数为 S，而用户 j 对于微博 v 的转发评论次数为 S_{jv}，则记边 $j\rightarrow v$ 的权重因子为 $\frac{S_{jv}}{S}$。

3. 结合1、2中阐述的网络关系，同时考虑用户之间的互粉关系以及用户与微博的转发评论关系，将二者融合成一个网络超图，即用户-微博-用户模型图。模型中既存在用户-用户之间的边，也存在用户-微博、微博-用户之间的边，权重仍按照上述方法计算，利用传播模型得到稳定的用户影响力值。

2.2 微博用户影响力的计算

本文采用加权PageRank算法计算用户的影响力。节点影响力的传播按照边的权重分配，而不是传统的均值划分。影响力计算算法如下：

第1步　根据用户-用户模型，利用影响力传播分析用户影响力排名，得到用户 i 的排名为 UU_i；

第2步　根据用户-微博模型，利用影响力传播分析用户影响力排名，得到用户 i 的排名为 UV_i；

第3步　计算用户影响力真实性系数即 $T_i=1-\frac{(UU_i-UV_i)^2}{N^2}$，其中 N 为用户节点个数；

第4步　根据用户-微博-用户模型，利用影响力传播分析用户影响力，计算得到用户 i 的影响力值为 pr_i，则用户最终的影响力值为 pr_i*T_i。

本文影响力计算算法并没有直接用用户-用户模型。用户-微博模型进行影响力的计算，一方面是因为仅仅根据用户关系衡量用户影响力结果往往不准确，由于微博中存在着大量的“僵尸用户”，这些用户平常不活跃，只是大量地关注其它用户，使其影响力得到提高；另一方面，用户-微博模型中同样存在着问题，微博中往往隐藏着大量的“水军用户”，这批用户大量地转发别人的微博内容，本身很少发布微博，同样会对用户的真实影响力带来偏差。影响力计算算法的第1、2两步中，分别计算这两个不同模型中用户的影响力排名，分析用户影响力排名的偏差量，用来衡量用户影响力真实性。若通过两种模型计算得到用户影响力排名比较一致，则可以说明用户的影响力排名比较真实，否则说明用户的影响力偏差比较大，即影响力结果比较可疑。因此，设定一个衡量用户影响力真实性的系数，该系数刻画了通过前两种模型计算的影响力的差别，差别越大，即用户的影响力可信度越低，所以该系数越小就好。最后，采用用户-微博-用户模型计算用户影响力值，该模型综合考虑了用户与用户之间的关系以及用户与微博之间的关系，并利用影响力真实性系数为该影响力值加权来得到用户 i 的最终影响力大小 $pr_i*(1-\frac{(UU_i-UV_i)^2}{N^2})$。通过真实性系数的加权，用户的影响力得到修正，这在一定程度上消除了“僵尸用户”和“水军用户”对结果的影响。

3 实验结果与分析

3.1 用户领域划分实验结果

本文采用基于微博内容和用户关系两种方法实现用户领域划分。首先单独利用两种方法进行领域划分，得到划分的结果，然后对结果加权合并，得到最终的领域划分结果。按照

新浪微博自身的用户领域划分体系，本文主要将用户划分为9个类别，分别为体育、政治、娱乐、教育、汽车、旅游、IT、游戏、生活。另外，由于用户领域划分采用的是无监督的聚类方法，因此，为了评估K-means聚类结果的准确率以及识别各个领域的类别信息，实验首先是人工采集标注了一批用户数据集，总共标记546个用户，标注结果如表1所列。

表1　用户标注数据集

类别	政治	体育	娱乐	教育	汽车	旅游	IT	游戏	生活
用户数	19	144	216	10	16	17	67	18	39

1. 基于微博内容的用户领域划分实验结果

对于收集到的150万个微博用户数据，由于实验数据规模较大，因此采用MapReduce并行计算模型对数据进行处理。根据用户的微博内容进行领域划分，标注数据划分结果如表2所列。

表2　标注数据集用户内容领域划分结果

	政治	体育	娱乐	教育	汽车	旅游	IT	游戏	生活	
聚类1	0	8	4	1	10	1	2	3	1	汽车
聚类2	0	112	16	0	1	2	1	0	2	体育
聚类3	2	3	5	5	0	0	7	0	0	教育
聚类4	4	2	27	2	0	3	3	0	20	生活
聚类5	11	0	4	0	1	0	1	0	0	政治
聚类6	0	2	5	0	0	8	0	0	11	旅游
聚类7	2	12	139	0	2	2	0	2	1	娱乐
聚类8	0	4	6	0	0	0	8	9	1	游戏
聚类9	0	1	10	2	2	1	45	4	3	IT

根据表2中的划分结果，标注数据中利用微博内容划分用户领域的准确率为65.8%。另外，为了验证微博新词发现和特征扩展是否能够提高微博用户领域划分的准确性，本文另外还进行了一组不加新词发现和特征扩展的对照实验，结果准确率为58.2%。由此可知，利用微博新词发现和特征扩展的方法确实可以提高微博用户领域划分结果的准确率。利用划分方法对整个微博用户数据进行划分，得到基于内容的微博用户领域划分结果，如图1所示。

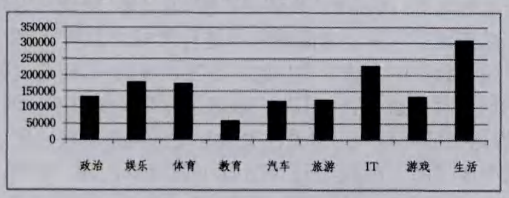


图1　基于微博内容的用户领域划分结果

2. 基于用户关注关系的用户领域划分实验结果

同样基于标注数据集，利用用户的关注关系对微博用户划分领域，实验结果如表3所列。

表3　标注数据集用户关注领域划分结果

	政治	体育	娱乐	教育	汽车	旅游	IT	游戏	生活	
聚类1	1	1	2	1	13	1	1	1	0	汽车
聚类2	0	120	3	1	1	0	3	0	2	体育
聚类3	2	1	2	4	0	1	0	1	3	教育
聚类4	4	5	4	0	0	4	1	2	26	生活
聚类5	10	0	3	0	0	0	0	0	0	政治
聚类6	0	3	6	2	0	6	4	1	4	旅游
聚类7	2	11	191	0	1	2	0	3	2	娱乐
聚类8	0	3	3	2	1	3	4	8	1	游戏
聚类9	0	0	2	0	0	0	54	2	1	IT

按照微博用户关注关系划分用户领域在标注数据集中结果的准确率为79.1%，根据用户关注划分微博用户领域的方法，筛选所有微博用户粉丝数大于5万的热门微博用户以及粉丝数大于500的核心微博用户共计25241人，根据用户关注列表共现划分用户领域，其他普通用户按照关注核心用户数量划分领域，划分结果如图2所示。

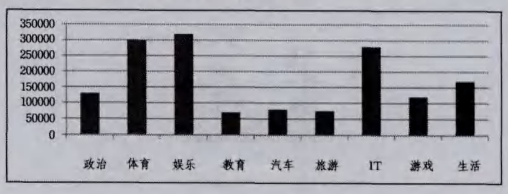


图2　基于用户关注的用户领域划分结果

3. 两种方法用户领域划分结果的合并

上述两种用户领域划分的方法均利用K-means聚类得到用户领域的结果，根据微博用户与每个领域中心相似度值归一化作为微博用户在各个领域中的权重值。将两种不同领域划分方法得到的权重值加权计算，取权重最大的作为微博用户的所属领域，在标注数据集中的实验结果准确率如表4所列。

表4　两种领域划分结果加权

按微博内容	按关注关系	准确率
0.1	0.9	82.23%
0.2	0.8	84.07%
0.3	0.7	82.23%
0.4	0.6	79.30%
0.5	0.5	76.74%
0.6	0.4	78.02%
0.7	0.3	72.89%
0.8	0.2	67.77%
0.9	0.1	66.67%

从结果可以看出，利用加权方法对标注数据集进行领域划分，基于微博内容划分权重0.2、基于关注关系划分权重0.8时加权得到的结果准确率最高。利用加权方式对微博用户数据集进行领域划分，最终得到划分之后的结果，对其进行影响力的分析。

3.2　微博用户影响力分析实验

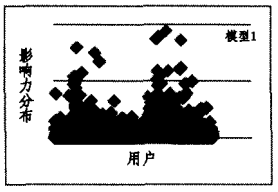
利用上述3种模型对完成领域划分的用户进行影响力分析，根据PageRank转播模型最终得到稳定的用户影响力转播状态，对结果进行分析统计。由于目前对微博用户的影响力分析中没有公共的评价数据集，因此，如何评价影响力结果是比较困难的。

各个领域中带有实名认证的用户往往较受关注，并且自身的影响力较大，利用这些实名认证标记的微博用户可以比较清晰地判断结果的好坏；另外，影响力较高的用户依据其不断传播的影响力，可以引起更多用户的关注，一段时间中用户的粉丝数量变化率也可作为用户影响力分析结果的评价指标。

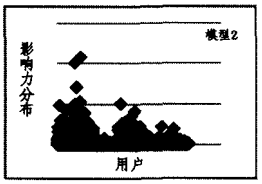
本文实验中首先利用上述3种模型构建用户影响力网络传播模型，在娱乐领域实验数据中得到的结果如图3所示。

在图3中，图(a)是利用用户-用户模型建模得到的用户影响力结果，图(b)是根据用户-微博模型建模得到的用户影响力分布，图(c)是利用用户-微博-用户模型得到的用户影响力结果。从3幅图中的结果可以看出，用户之间的相互关注关系对用户影响力计算贡献较大，而利用用户为微博的转发

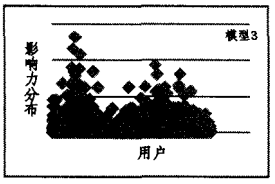
评论则贡献较小；图(c)根据图(a)、图(b)两幅图的模型合并得到最终的结果，与预期较符合，可以对前两种建模结果起到平滑优化的效果。另外，分析结果也可得到，实验语料中，娱乐领域用户影响力分布大都较平稳，少量影响力较高的用户则分布比较明显。



(a)用户-用户模型



(b)用户-微博模型



(c)用户-微博-用户模型

图 3　用户影响力分析建模结果

实验中采用 3 种建模结果加权的方式得到用户影响力结果，计算公式为 $pr_i * (1-\frac{(UU_i-UV_i)^2}{N^2})$，在结果中选取影响力排名较高的 15 位用户，结果如表 5 所列。

表 5　娱乐领域本文算法最终影响力前 15 的用户

排名	用户 ID	用户昵称	是否认证	粉丝数	影响力值
1	1195230310	姚晨	1	60744703	0.00132471
2	1195230310	何炅	1	41562686	0.0011603
3	1191258123	韩寒	1	33068597	0.001074256
4	1618051664	头条新闻	1	27857088	9.27E−04
5	1947267610	电影梦工厂	1	1016831	8.94E−04
6	1723141197	毕福剑	1	709068	8.73E−04
7	1265057942	芙蓉姐姐	1	4225540	8.40E−04
8	1665074831	谢楠	1	3478251	8.38E−04
9	1781387491	iphone 客户端	1	21121627	8.02E−04
10	1642909335	微博小秘书	1	82475363	8.00E−04
11	2481092927	韩饭桶	1	315580	7.87E−04
12	1220291284	作业本	1	7480872	7.10E−04
13	1063890337	洪峰	1	2039322	6.81E−04
14	1195031270	郑渊洁	1	6302558	6.62E−04
15	1191220232	高晓松	1	35451093	6.60E−04

另外本文利用微博用户节点的入度数(即用户的粉丝数)得到的影响力排名前 15 的微博用户如表 6 所列。

表 6　娱乐领域按粉丝数排名前 15 的用户

排名	用户 ID	用户昵称	是否认证	粉丝数
1	1642909335	微博小秘书	1	82475363
2	1195230310	姚晨	1	60744703
3	1197161814	李开复	1	51647465
4	1195230310	何炅	1	41562686
5	1191220232	高晓松	1	35451093
6	3125046087	刘烨	1	35046133
7	1191258123	韩寒	1	33068597
8	1618051664	头条新闻	1	27857065
9	1781387491	iphone 客户端	1	21121627
10	1914100420	稀土部队	0	17985865
11	1220291284	作业本	0	7480872
12	1761179351	留几手	0	6378731
13	3880051096	王诗龄 Angela	1	5366472
14	1265057942	芙蓉姐姐	1	4225540
15	1895964183	一起神回复	0	3661719

从认证用户的角度，在收集数据中，按节点入度排序时影响力排名前 15 的娱乐领域中的人物有 4 名用户是非认证用户，而在按本文方法计算的前 15 名用户均为认证用户。另外，考虑到一段时间内，微博用户粉丝数的变化情况可以反映用户影响力大小，本文实验中统计出在 120 天的时间跨度中微博用户粉丝数的变化率，本文方法排名计算得到的娱乐领域中前 15 名的用户粉丝的总变化率为 28.21%，而按入度方式粉丝的变化率为 22.13%。由此可见，本文计算出的排名靠前的用户在一段时间后，粉丝的增长量高于按入度计算的方法，在一定程度上说明用户的影响力排名更加科学。

本文最后统计各个领域中用户粉丝数变化率情况，如表 7 所列。

表 7　认证用户和变化率的分布

		前 1%	前 10%	后 90%
娱乐	认证用户	76.33%	11.71%	0.26%
	变化率	18.41%	3.08%	/
IT	认证用户	78.91%	12.63%	0.05%
	变化率	15.17%	2.96%	/
体育	认证用户	70.39%	13.87%	0.81%
	变化率	16.70%	0.93%	/
汽车	认证用户	58.23%	12.43%	4.87%
	变化率	20.18%	4.86%	/
游戏	认证用户	51.41%	19.28%	6.78%
	变化率	10.43%	2.90%	/
教育	认证用户	45.87%	12.66%	2.87%
	变化率	7.91%	2.89%	/
生活	认证用户	36.76%	27.68%	3.23%
	变化率	4.55%	3.24%	/
政治	认证用户	54.76%	11.43%	0.92%
	变化率	6.90%	0.29%	/
旅游	认证用户	48.04%	8.91%	2.30%
	变化率	10.34%	1.01%	/

由表 7 可知，影响力排名前 1%的认证用户明显较多，而在影响力排名靠后的位置，认证用户的比例明显减少，这一点在体育类、IT 类、娱乐类人物中尤为明显，应该与该领域知名的人物较多有关，用户真正的关注点在这些明星人物身上，而对于其他类别，认证用户大多是机构、网站等，数量相对较少。

结束语　本文针对微博用户进行影响力分析，首先利用用户领域划分的方法对用户微博进行领域划分，在此之中，利用新词发现算法、特征词语扩展等方法提高领域划分的准确率，根据微博内容和用户关注关系来划分用户领域。之后，利用文中所述的 3 种模型计算出各个领域中的用户影响力大小，并且对结果进行了分析。另外，本文在验证结果准备率中采用人工标注的数据集，由于人工标注的数据集中数据分布存在偏差，可能导致在某些领域中结果较好而在其它领域中的结果有待提高。因此如何采用更为有效的验证方法是今后研究的目标。

参 考 文 献

[1] Wu Shao-mei, Hofman J M. Who says what to whom on Twitter [C]// Proc. of the 20th Int. Conf. on World Wide Web (2011). Hyderabad, India, 2011: 705-714

[2] Page L, Brin S, Motwani R, et al. The PageRank citation ranking: bringing order to the Web[OL]. http://ilpubs.stanford.edu:8090/4221

[3] He X, Li Y, Fan C. Web-based links and authoritative content Pagerank Improvement[C]// 2010 International Conference on E-Business and E-Government (ICEE). IEEE, 2010: 5016-5019

(下转第 66 页)

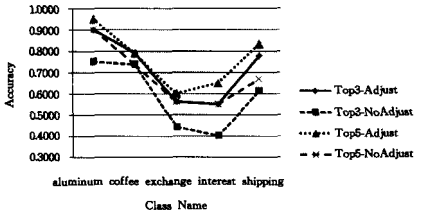


图 3　Top3、Top5 主题挖掘比较示意图

从图 3 可以看出，选择前 3 个概念或前 5 个概念作为最终主题的两种情况下使用标签和位置对特征向量进行权重修正(Adjust)比不使用标签和位置进行权重修正(NoAdjust)的准确率高；且同样进行权重修正时，其准确率差异在主题概念选择较少时表现更为突出。

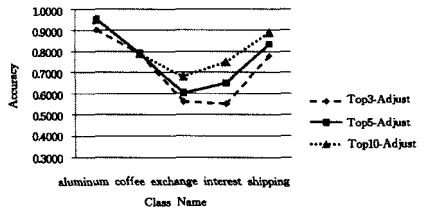


图 4　Top3、Top5 和 Top10 主题挖掘比较示意图

从图 4 可以看出，随着最终所选择主题概念数的增加，主题挖掘正确率增加，且在选择前 10 个概念作为主题概念的情况下，各类别准确率都在 0.68 以上，最高达到 0.95。

对于 n 值的选择，通过实验可以看到：随着 n 值的增大，主题概念挖掘的准确率也提高。但是选择的 n 个概念中非主题概念的个数也随之提升，且能表示网页主题的概念数一般不会太多，所以 n 值应该在一个较小且包含主题概念较多的范围内。从图 4 可以看出，在 n 取 5 时，算法准确率已经较高，因此，最终的 n 值可以取 5，并依据具体数据进行调整，以提高主题挖掘的准确率，并减少噪声出现的次数。

结束语　本文提出了以概念代替单词构建文本特征向量并使用标签和位置对待测网页中的候选主题概念进行权重修正得到最终主题的方法，分别给出了标签和位置权重计算公式，将标签分为可见标签和不可见标签两种情况，充分利用了网页的 HTML 代码蕴含的信息。通过实验验证了根据标签和位置修正网页文本特征的必要性，且主题挖掘具有较高的准确率。对于中文网页，可以使用中科院提供的分词和词性标注工具对中文网页正文进行分词和词性标注，并基于 Hownet 进行词义消歧。但受到词义消歧效率的限制，本文进行实验的数据集较小，尚需改进。

参 考 文 献

[1] Jayabharathy J, Kanmani S, Parveen A A. Document Clustering and Topic Discovery based on Semantic Similarity in Scientific Literature[C]// 2011 IEEE 3rd International Conference on Communication Software and Networks (ICCSN). 2011: 425-429

[2] Uluhan E, Badur B. Development of a Framework for Sub-Topic Discovery from the Web[C]// PICMET 2008 Proceedings. July 2008: 878-888

[3] Shi Jing, Li Wan-long. Topic Discovery Based on LDA Model with Fast Gibbs Samping[C]// 2009 International Conference on Artificial Intelligence and Computational Intelligence. 2009: 91-95

[4] Ding W, Rohban M H, Ishwar P, et al. Topic Discovery through Data Dependent and Random Projections[C]// International Conference on Machine Learning (ICML'13). 2013: 471-479

[5] Yang Yun, Wu Ya-nan. Content-based topic discovery of high-impact model[C]// 2010 2nd International Conference on Computer Engineering and Technology. 2010

[6] 王琦，唐世渭，杨冬青，等. 基于 DOM 的网页主题信息自动提取[J]. 计算机研究与发展，2004，41(10)：1756-1792

[7] Yamaguchi Y, Amagasa T, Kitagawa H. Tag-based User Topic Discovery using Twitter Lists[C]// 2011 International Conference on Advances in Social Networks Analysis and Mining. 2011: 13-20

[8] Cheng L. Unsupervised topic discovery by anomaly detection [D]. Monterey, California: Naval Postgraduate School, 2013

[9] Pedersen T, Banerjee S, Patwardhan S. Maximizing semantic relatedness to perform word sense disambiguation[J/OL]. http://www.patwardhans.net/papers/pedersenBP05.pdf

[10] Naskar S K, Bandyopadhyay S. Word sense disambiguation using extended wordnet[C]// Proceedings of the International Conference on Computing: Theory and Applications (ICCTA'07). 2007: 446-450

[11] Naskar S K, Bandyopadhyay S. JU-SKNSB: extended WordNet based WSD on the English all-words task at SemEval-1[C]// Proceedings of the 4th International Workshop on Semantic Evaluations. Association for Computational Linguistics. 2007: 203-206

[12] Shen Wan, Angryk R A. Measuring semantic similarity using wordnet-based context vectors[C]// IEEE International Conference on Systems, Man and Cybernetics, 2007(ISIC). 2007: 908-913

(上接第 46 页)

[4] Zhang J, Ackerman M, Adamic L. Expertise networks in online communities: structure and algorithms[C]// Proc. of the 16th Int. Conf. on World Wide Web (2007). 2007: 221-230

[5] Tang J, Sun J, Wang C, et al. Social Influence Analysis in Large-scale Networks[C]// Proc. of the 15th Int. Conf. on Knowledge Discovery and Data Mining (SIGKDD 2009). Paris, France, 2009: 807-816

[6] 田军伟. 基于社会网络的用户兴趣模型研究[D]. 成都：电子科技大学，2010

[7] Sharifi B, Hutton M-A, Kalita J K. Experiments in Microblog Summarization[C]// IEEE Second International Conference on Social Computing. 2010: 49-56

[8] Yang Y. Expert network: Effective and efficient learning from human decisions in text categorization and retrieval[C]// Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (1994). 1994: 13-22

[9] Cortes C, Vapnik V. Support vector networks [J]. Machine Learning, 1995, 20: 273-297

[10] Feng Hao-di, Chen Kang, Deng Xiao-tie, et al. Accessor Variety Criteria for Chinese Word Extraction[J]. Computational Linguistics, 2004, 30(1): 75-93

[11] 沈崇玮. 基于微博数据的用户影响力分析研究[D]. 北京：北京邮电大学，2013