图书情报工作
第61卷 第9期 2017年5月

融入情感差异和用户兴趣的微博转发预测*

唐晓波[1,2] 罗颖利[1]
[1]武汉大学信息管理学院 武汉 430072 [2]武汉大学信息系统研究中心 武汉 430072

摘要：[目的/意义]微博转发是实现微博信息传播的重要方式，对用户转发行为进行研究可以更好地理解微博信息传播机制，对热点话题检测、舆情监控、微博营销等具有重要意义。针对以往研究中用户兴趣表示不够全面准确以及未考虑情感差异对用户转发行为的影响，提出一个融入情感差异和用户兴趣的微博转发预测模型。[方法/过程]该模型首先从维基百科中提取概念语义关系构建维基知识库，将其作为语义知识源对微博文本进行语义扩展，解决语义稀疏问题；对语义扩展后的用户历史微博进行聚类，提取用户兴趣主题和主题对用户的影响力；然后计算微博中各类情感的情感强度，提取情感差异特征；最后结合用户行为特征、用户交互特征、微博特征、用户兴趣特征和情感差异特征，运用SVM实现微博转发预测。[结果/结论]在新浪微博真实数据集上进行实验，验证了所提模型的有效性。

关键词：微博转发 用户兴趣 语义扩展 情感差异 情感分析
分类号：TP18
DOI：10.13266/j.issn.0252-3116.2017.09.013

1 引言

近年来，互联网的普及与在线社交网络的快速发展，吸引越来越多的人利用社交网络分享和传播观点。微博作为信息获取、分享和传播的开放信息平台，已经成为人们传递信息、表达舆论、抒发情绪的重要媒体。不同于基于内容的Web信息传播，微博以人为中心，能够通过用户关系以及转发行为进行信息的快速扩散，在短时间内形成极大影响和关注[1]。对微博用户的转发行为进行研究，一方面能够更好地理解用户行为，挖掘用户兴趣，为用户兴趣建模、微博推荐、好友推荐等研究提供理论基础；另一方面能够较为准确地预测出微博的传播范围和发展趋势，对热点话题检测、突发事件预警、舆情监控、微博营销等具有重要价值。

个体的转发行为影响着微博信息整体的传播趋势，从微观角度研究和预测用户的转发行为具有重要意义。用户转发微博的行为受用户兴趣所驱动，目前关于微博转发预测的研究往往使用词频统计方法从用户历史微博中提取高频词，无法全面表示用户兴趣偏好。微博内容包含的情感信息也是影响用户转发行为的重要因素，情感差异程度大的微博往往携带了更多的衍生信息，更容易得到转发[2]。本文以此为出发点，提出一个融入情感差异和用户兴趣的微博转发预测模型。该模型首先利用维基百科构建维基知识库，并利用构建好的维基知识库对微博文本进行语义扩展，挖掘潜在语义信息，从而准确全面提取用户兴趣；然后利用情感分析方法计算待预测微博的情感差异程度，考虑情感差异对用户转发行为的影响；最后对分析影响用户转发行为的各类因素并建立特征建模，运用SVM对特征模型进行分类进而实现微博转发预测。

2 相关研究

微博转发作为社交网络中信息传播的重要方式和机制，引起了研究者们的广泛关注，国内外对微博转发的研究已经取得了一些成果。D. Boyd等[3]对Twitter的转发功能做了细致的分析，探讨了Twitter上人们的转发方式和转发动机。B. Suh等[4]以Twitter为研究对象，分析了影响转发的各种因素，结果表明，微博中是否含有URL和hashtag标签对转发影响较大。X. Zhao等[5]发现，相比于普通微博，含有图片、视频等多

* 本文系国家自然科学基金项目"基于文本和Web语义分析的智能咨询服务研究"（项目编号：71673209）研究成果之一。
作者简介：唐晓波（ORCID：0000-0001-5885-45090），教授，博士生导师；罗颖利（ORCID：0000-0002-4149-4831），硕士研究生，E-mail：2579764241@qq.com。
收稿日期：2017-01-09 修回日期：2017-04-08 本文起止页码：102-110 本文责任编辑：徐健

媒体内容的微博转发量更大。J. Bian 等[6]分析了微博流行性对用户转发行为的影响，并在此基础上提出了一种传播模型对转发行为进行预测。张玢玢等[7]以企业官方微博为研究对象，研究微博内容的文本特征和形式化特征对微博转发的影响及影响规律。张旸等[8]运用信息增益方法对微博上不同特征的重要性进行分析，提出基于特征加权的模型对微博转发进行预测。曹玖新等[9]对各种可能影响用户转发行为的因素并进行统计分析，综合用户属性特征、微博内容特征和社交关系特征，预测用户对给定微博的转发行为。刘玮等[10]基于用户转发率、交互频率等用户行为特征，融合微博特征和用户兴趣特征，运用分类模型进行转发预测。这些研究主要使用用户静态属性、用户交互关系和微博本身特征来预测微博是否会被转发，没有充分考虑用户的个体差异对转发行为的影响。

事实上，用户转发微博的行为受用户兴趣所驱动，当用户看到一条微博时，往往会根据个人兴趣和理解对微博价值和新颖性进行判断，然后决定是否进行转发[11]。有效提取用户兴趣对提高微博转发预测准确率具有重要作用。谢婧等[12]利用互信息理论从转发用户群的微博内容中提取特征，分析用户内容与特征之间的相关程度，预测用户是否会转发给定主题的微博。吴凯等[13]从用户历史微博中提取高频词集合，利用 Jaccard 相似度计算微博内容与用户兴趣之间的相似关系，结合用户特征和文本特征，利用逻辑回归模型预测个体的转发行为。这些研究主要从用户历史微博中提取高频词作为用户兴趣，忽略了词语之间的语义关系和低频词，无法全面表示用户兴趣。李志清[14]在用户历史微博文本集上使用 LDA 抽取用户兴趣主题，结合用户特征和微博特征建立转发预测模型。但微博属于短文本，直接运用 LDA 模型效果并不理想。

微博成为人们传播信息、表达舆论、抒发情绪的重要媒体，微博内容的情感信息表达了用户对事物的态度，对微博转发具有重要影响。S. Stieglitz 等[15]发现微博内容的情感与其被转发情况存在相关性，与不带情感倾向的微博相比，包含情感信息的微博更容易得到转发。A. Kanavos 等[16]构建微博情绪模型，基于微博内容的情感倾向预测微博传播的广度和深度。N. Naveed 等[17]分析了影响微博转发的诸多因素，发现情感特征是影响用户转发行为的重要因素，带有消极情感倾向的微博转发量更高。R. Pfitzner 等[2]研究表明，情感差异性是影响微博转发的重要因素，情感差异大的微博往往携带更多的衍生信息，更容易引起用户转发。这些研究表明微博内容的情感与其被转发情况存在相关性，但仅局限于对转发数量的研究，尚未探索情感强度和情感差异在微博转发预测中的作用。

综上所述，微博转发预测研究还存在以下不足：①没有准确全面表示用户兴趣，影响转发预测准确率；②尚未考虑微博内容的情感强度和情感差异对转发预测的影响。因此，本文主要针对这两点不足提出一个融入情感差异和用户兴趣的微博转发预测模型。利用维基知识库对微博文本进行语义扩展，挖掘潜在语义信息，对扩展后的微博文本进行聚类得到用户兴趣主题及主题对用户的影响力，全面表示用户兴趣偏好；利用情感分析方法计算微博中各类情感的情感强度，提取情感差异特征，探索情感差异特征对微博转发预测的作用；最后对用户行为特征、用户交互特征、微博特征、用户兴趣特征和情感差异特征进行特征建模，运用 SVM 对特征模型进行分类进而实现微博转发预测。

3 融入情感差异和用户兴趣的微博转发预测模型

本文提出融入情感差异和用户兴趣的微博转发预测模型。该模型包括 4 个模块：维基知识库构建模块、微博文本语义扩展模块、转发特征分析与提取模块、微博转发预测模块（见图 1）。

3.1 维基知识库构建模块

很多学者使用知网和同义词词林来解决语义稀疏问题，但并不适用于微博文本，因为微博文本涉及领域广、词语更新快，而知网和同义词词林的信息更新速度慢、知识覆盖率低。维基百科知识高度结构化、覆盖面广且更新速度快。为此，本文利用维基百科构建维基知识库，作为语义知识源对微博文本进行语义扩展。维基知识库构建主要分为 4 个部分：维基数据预处理、相关概念抽取、概念间的语义关系量化以及维基知识库表示。

3.1.1 维基数据预处理 从维基百科网站[18]下载最新的维基百科中文 XML 语料集，主要包含 3 个文件，如表 1 所示：

表 1 本文下载的中文维基百科文件

文件名	文件描述
pages-articles. xml. bz2	概念解决页面
pagelinks. sql. gz	页面链接关系
categorylinks. sql. gz	类别链接关系

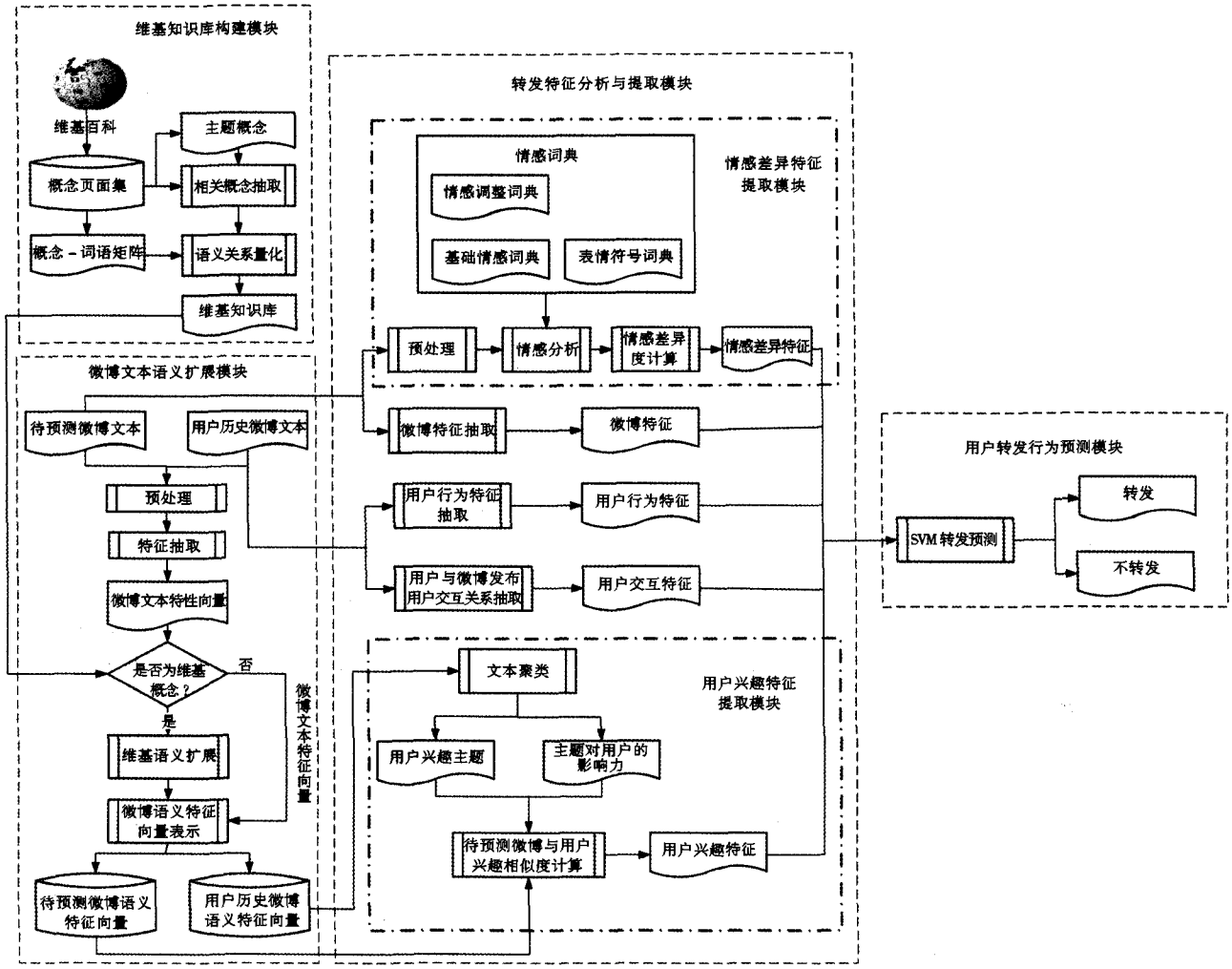

图1 融入情感差异和用户兴趣的微博转发预测模型

本文使用JWPL(Java Wikipedia Library)工具[19]处理维基语料。利用下载的JWPL DataMachine jar包解析表1中的3个文件,共产生11个txt文件。由于中文维基百科中存在简体和繁体两种文字形式,对生成的11个txt文件进行繁简转换处理,去掉特殊符号、年代名词和冗余信息后导入MySQL数据库。

3.1.2 相关概念抽取 维基百科概念页面间的内部链接关系是进行相关概念抽取的最佳语义资源。如果概念 C_t 解释页面的入链和出链都包含了概念 C_i,说明概念 C_t 和 C_i 在语义上具有很强的相关度。与双向链接关系相比,概念页面间的单向链接关系相对较弱,如果将这些语义关联较弱的概念作为相关概念进行扩展,将增加维基知识库的冗余性。因此,本文运用*JWPL*工具抽取出与概念 C_t 具有双向链接关系的概念,构建 C_t 的相关概念集合,表示为 $C_t(C_1, C_2, \cdots\cdots, C_n)$,其中 n 为与概念 C_t 具有双向链接关系的相关概念的个数。

3.1.3 概念间的语义关系量化 对上述抽取出的概念间的语义关系进行量化,明确各概念与其相关概念间的语义关联强度。本文借鉴显性语义分析方法(explicit semantic analysis, ESA)[20]进行概念间语义关联度计算。维基百科页面是对概念的描述和解释,对维基概念页面文本进行分词和去停用词,提取特征词并利用TF-IDF计算其权重,形成维基概念-词语矩阵,矩阵中的行代表维基概念,每个维基概念由出现在页面文本中的词向量表示。概念 C_t 与其相关概念 C_i 之间的语义关联度可以利用向量间的余弦相似度来计算。

$$R(C_t, C_i) = \cos(C_t, C_i) = \frac{C_t \cdot C_i}{\| C_t \| * \| C_i \|}$$

公式(1)

其中,C_t、C_i 分别为概念 C_t、C_i 的词向量,$\| C_t \|$、$\| C_i \|$ 表示概念向量的模,$R(C_t, C_i)$ 为概念 C_t、C_i 间的语义关联度。$R(C_t, C_i)$ 越大,表示概念间的语义相

关性越强，对语义信息的补充能力也就越强。

3.1.4 维基知识库表示 经过相关概念抽取和概念间的语义关系量化后，概念 C_t 的相关概念集合可以表示为 $C_t((C_1,R_1),(C_2,R_2),\cdots\cdots,(C_n,R_n))$，其中 C_i 表示与概念 C_t 具有双向链接关系的相关概念，表示第 i 个相关概念 C_i 与概念 C_t 之间的语义关联度，由公式(1)求得。综上各步骤，可以构建维基知识库，知识库中包括概念和相关概念集合及其语义关联度，如表2所示：

表2 维基知识库(部分)

概念	相关概念及语义关联度
苹果公司	(iPhone,0.371 1),(IOS,0.245 3),(Macbook,0.242 5)
互联网	(万维网,0.273 6),(电子商务 0.1728),(ARPA 网,0.147 4)
人工智能	(机器学习,0.323 4),(人工神经网络,0.321 5),(专家系统,0.283 7)
笔记本电脑	(个人电脑,0.529 1),(触控板,0.253 8),(键盘,0.167 3)

3.2 微博文本语义扩展模块

由于微博属于短文本，所含特征少、文本特征稀疏，因此有必要对其进行语义扩展，提高微博文本语义描述能力。首先将预处理后的微博文本以向量的形式进行表示，然后使用3.1中构建的维基知识库对微博文本特征向量进行语义扩展，得到微博文本语义特征向量。

3.2.1 微博文本特征向量表示 微博文本具有随意、不规范等特点，因此有必要在实验前对其进行预处理，主要工作包括去噪处理、分词、去停用词和标点符号等。在中文语料中，名词和动词对主题表达作用最大，因此本文在预处理之后对微博文本进行词性过滤，提取动词和名词，然后利用 TF-IDF 算法提取关键词，得到微博文本特征向量 $d=((t_1,w_1),(t_2,w_2),\cdots\cdots,(t_n,w_n))$，其中 t_i、w_i 分别表示特征词及其权重。

3.2.2 维基语义扩展及语义特征向量表示 将微博文本向量中的特征词与维基知识库进行匹配，根据匹配到的概念取出其相关概念，作为扩展特征词加入原始微博向量中，以此完善微博文本的语义信息。利用维基百科的重定向和消歧机制，解决自然语言中存在的同义词和多义词问题。详细过程如下所述：

步骤一：特征-概念匹配。将微博文本特征向量中的所有特征词 t_i 依次与维基知识库中的概念进行匹配，存在以下3种情况：

(1)匹配成功，且维基知识库中存在唯一的概念 C_t 与特征词 t_i 相匹配，则取出概念 C_t 的相关概念集合 $C_t((C_1,R_1),(C_2,R_2),\cdots\cdots,(C_n,R_n))$。当与特征词 ti 相匹配的概念 Ct 在维基百科中对应的是重定向页面时，根据重定向机制找到的同义概念，并取出同义概念的相关概念集合。

(2)匹配成功，且维基知识库中存在多个概念与特征词 t_i 相匹配，即特征词具有多个含义，因此需要进行词义消歧处理，具体处理方法如下：首先，找到与特征词 t_i 相匹配的所有概念及其概念-词语特征向量；然后，将特征词 t_i 所在的微博文本特征向量中除 t_i 以外的其他特征词作为 t_i 的语境信息，组成向量 D；最后，利用余弦相似度将向量 D 与每个概念-词语特征向量进行语义比较，将相似度最大的概念作为与特征词 t_i 相匹配的概念，并取出该概念的相关概念集合。

(3)匹配失败，即维基知识库中不存在与特征词 ti 相匹配的概念，则不做扩展，直接跳到步骤四。

步骤二：微博文本特征向量扩展。根据匹配到的概念 C_t，取出相关概念集合中相似度最高的3个概念，作为扩展特征词加入到微博文本特征向量中。

步骤三：扩展特征词权重计算。扩展特征词的权重既与相应概念在原始微博向量中的权重有关，又与概念间的语义关联度有关。因此，本文采用以下公式计算扩展特征词权重：

$$w_{ij}=w_i*R_v \qquad \text{公式(2)}$$

其中，w_i 表示被扩展特征词 t_i 在原始微博文本特征向量中的权重，R_{ij} 表示概念 C_t 与其相关概念 C_j 之间的语义关联度。

步骤四：微博文本语义特征向量表示。经过语义扩展后，对重复出现的特征词进行合并处理，并对其权重求和。经过上述处理，得到扩展后的微博文本语义特征向量 $d=((t_1,w_1),(t_2,w_2),\cdots\cdots(c_{11},w_{11}),(c_{12},w_{12}),(c_{13},w_{13}),\cdots\cdots,(t_n,w_n))$。

3.3 转发特征分析与提取模块

转发特征的选择将在很大程度上影响转发预测的准确率，本模块分析用户转发行为的影响因素，提取转发特征，为转发预测做准备。本文选取的特征主要有以下几类：用户行为特征、用户交互特征、微博特征、用户兴趣特征、情感差异特征。

3.3.1 用户行为特征 用户的行为特征对用户转发行为具有重要影响，本文中的用户行为特征主要指用户转发活跃度，即用户历史微博中转发微博所占的比例。用户转发活跃度能够代表用户在微博网络中是倾向于转发微博还是原创行为，用户的转发活跃度越高，转发微博的概率越大。本文按如下公式计算用户转发活跃度：

$$f_n = \frac{\text{时间 } t \text{ 内用户转发的微博数量}}{\text{时间 } t \text{ 内用户发布的微博数量}} \quad \text{公式(3)}$$

3.3.2 用户交互特征 用户交互特征指接收用户与发布用户之间的交互强度。研究表明,用户更倾向于转发与自己关系紧密的用户的微博[21]。用户 u 与用户 v 的交互强度越大,说明用户 u 转发用户 v 的微博的概率越大。研究接收用户与发布用户之间的交互强度,能够找出对用户转发行为影响力大的发布用户,对微博转发预测具有重要作用。本文按如下公式计算用户之间的交互强度 f_{uv}:

$$f_{uv} = \frac{\text{用户 } u \text{ 转发用户 } v \text{ 的微博数量}}{\text{用户 } u \text{ 的转发微博数量}} \quad \text{公式(4)}$$

3.3.3 微博特征 微博信息内容的丰富性,如:是否包含图片、视频等,都会影响用户对微博的转发行为。因此,本文提取以下微博特征:是否包含 hashtag、是否包含 URL、是否包含图片、是否包含视频、是否包含@。

3.3.4 用户兴趣特征 用户转发微博的行为受用户兴趣所驱动,有效提取用户兴趣特征对提高微博转发预测准确率具有重要作用。本文中的用户兴趣特征包括待预测微博与用户兴趣的相似度及待预测微博所属主题对用户的影响力。微博内容与用户兴趣的吻合程度越高,所属主题对用户的影响力越大,用户转发该微博的可能性越大。本文按如下步骤提取用户兴趣特征:

步骤一:用户兴趣收集。用户发布的历史微博体现了用户兴趣,但用户兴趣具有时效性。因此,本文选择用户最近发布的100条微博作为用户兴趣提取的数据集,记为 D_u。

步骤二:用户兴趣主题及主题影响力计算。由于微博属于短文本,所含特征少,直接对其进行聚类效果不佳。因此,本文在预处理后利用3.2中的方法对用户历史微博集 Du 进行语义扩展,然后使用 K-means 算法对其进行聚类,将聚类后每个类别的特征词按 TF-IDF 值进行排序,取排名最高的10个词作为这一类别的主题特征词,得到用户的兴趣主题及主题向量。通常情况下,用户对主题的兴趣度决定主题对用户的影响力,用户对各个主题的兴趣度不同,主题对用户的影响力也就不同。可以通过用户的历史行为来衡量用户对主题的兴趣度,如果用户对某个主题的兴趣度高,则用户会发布或转发更多的关于该主题的微博。本文采用如下公式计算主题对用户的影响力:

$$Influence = \frac{N_i}{N} \quad \text{公式(5)}$$

其中,$Influence_i$ 表示主题 i 对用户的影响力,N 表示用户历史微博数量,N_i 表示用户历史微博中属于主题 i 的微博数量。

步骤三:用户兴趣特征提取。利用3.2中的方法对待预测的微博文本进行语义扩展,得到语义特征向量 d,利用余弦相似度计算语义特征向量 d 与用户所有兴趣主题的相似度,将相似度最大的主题作为待预测微博所属的主题,将待预测微博与所属主题的相似度作为其与用户兴趣的相似度,结合其所属主题的对用户的影响力,构成用户兴趣特征。

3.3.5 情感差异特征 情感差异(emotional divergence)指微博的归一化正向情感值和负向情感值的差异。情感差异程度大的微博往往包含更多的衍生信息,更容易引起用户的转发[2]。因此,本文将情感差异特征融入微博转发预测中。首先对微博进行情感分析,得到该条微博的正向情感值和负向情感值,然后按公式(9)计算微博的情感差异特征。具体过程描述如下:

步骤一:构建情感词典。情感词典的全面性严重影响着情感分析的准确性,本文使用基础情感词典的同时,构建微博表情符号词典和程度副词情感调整词典。

本文使用的基础情感词典是大连理工大学情感词汇本体库[22],共包含27 466个情感词,其中正面情感词11 229个,负面情感词10 782个,基本格式如表3所示:

表3 情感词汇本体库(部分)

词语	词性种类	词义数	词义序号	情感分类	强度	极性
博文广识	idiom	1	1	PH	5	1
骚乱	noun	1	1	NN	3	2
贪污	verb	1	1	NN	7	2
创新	verb	1	1	PH	5	1

另外,微博中含有丰富的表情符号,如"[威武]""[泪]"等,这些表情符号可以辅助情感分析。本文从新浪微博平台上抓取了84个最常见的表情符号,人工标记每个表情符号的情感强度和极性,建立了微博表情符号词典,如表4所示:

表4 微博表情符号词典(部分)

表情符号	强度	极性	表情符号	强度	极性
泪流满面	7	2	威武	5	1
怒	7	2	给力	5	1
伤心	5	2	赞	5	1
委屈	3	2	偷乐	3	1

程度词和否定词对情感计算具有重要影响。对常用的程度词，本文构建了程度词情感调整词典，共分为4种调整力度，分别为：1.6，1.3，1，0.7，如表5所示：

表5 情感调整词典(部分)

情感调整力度	程度词
1.6	非常 极其 绝对 极度 无比 极 万分 特别
1.3	厉害 真心 很 挺
1	无程度词
0.7	略微 稍微 有点儿

步骤二：计算正向情感值和负向情感值。对微博文本进行预处理，根据基础情感词典识别出微博文本中的正向情感词、负向情感词以及情感词对应的情感强度，根据情感调整词典识别程度词的调整力度，然后按公式(6)计算微博中每个情感词的情感得分；对于存在表情符号的微博，根据正则表达式提取微博文本中的表情符号，从构建的表情符号词典中匹配出其情感极性和情感强度；最后按公式(7)和公式(8)分别计算该条微博的正向情感值和负向情感值。

情感词的情感得分计算公式如下：

$$\text{Sentiment}(w_i) = s(w_i) * (-1)^n * str(adv) \quad \text{公式(6)}$$

其中，Sentiment 表示微博中情感词 w_i 的情感得分，$s(w_i)$ 表示情感词 wi 的情感强度，n 表示 w_i 附近的否定词数量，$str\ (adv)$ 表示 w_i 附近程度词的情感调整力度。

微博的正向情感值和负向情感值计算公式如下：

$$sentiment\text{-}P = \sum_{sentiment(w_i)>0} sentiment(w_i) + \sum_{s(e_j)>0} s(e_j) \quad \text{公式(7)}$$

$$\text{sentiment-N} = \sum_{sentiment(w_i)<0} sentiment(w_i) + \sum_{s(e_j)<0} s(e_j) \quad \text{公式(8)}$$

公式(7)和(8)中的 *Sentiment-P* 和 *Sentiment-N* 分别表示微博的正向情感值和负向情感值，是对微博中所有情感词和表情符号的正向情感得分和负向情感得分分别进行相加求和的结果。

步骤三：提取情感差异特征。由于存在程度副词对情感强度的调整，且每条微博中包含的情感词个数不同，不同微博的情感得分存在很大的差距。因此，本文对正向情感值 *Sentiment-P* 和负向情感值 *Sentiment-N* 分别进行归一化处理，然后利用公式(9)计算微博的情感差异特征 f_{ed}：

$$f_{ed} = \frac{p-n}{2} \quad \text{公式(9)}$$

其中，p 表示微博的归一化正向情感得分，n 表示微博的归一化负向情感得分，$e \in (0,1]$，$n \in [-1,0)$。

通过上述分析，本文共提取了影响用户转发行为的10个特征，如表6所示：

表6 转发模型特征分析

特征类别	特征序号	特征描述
用户行为特征	1	用户转发活跃度
用户交互特征	2	接收用户与发布用户之间的交互强度
微博特征	3	待预测微博是否包含 hashtag
	4	待预测微博是否包含 URL
	5	待预测微博是否包含图片
	6	待预测微博是否包含视频
	7	待预测微博是否包含@
用户兴趣特征	8	待预测微博与用户兴趣的相似度
	9	待预测微博所属主题对用户的影响力
情感差异特征	10	待预测微博的情感差异程度

3.4 微博转发预测模块

微博中的信息是通过关注网络进行传递的，微博转发预测是研究关注网络中的消息传播过程。如果用户 u 的关注用户 v 发布了一条微博 m，则 $y=f(u,v,m)$ 表示用户 u 在看到该微博后是否会产生转发行为，$y \in \{0,1\}$，$y=0$ 表示不转发，$y=1$ 表示转发。因此，可以将微博转发预测转化为二分类问题来处理。本文根据待预测微博和用户发布的历史微博，提取用户行为特征、用户交互特征、微博特征、用户兴趣特征和情感差异特征，运用台湾大学林志仁教授等开发的 LIBSVM 工具包[23]，使用径向基核函数，实现微博转发预测。

4 实验及结果分析

4.1 实验数据

本文利用网络爬虫工具八爪鱼[24]从新浪微博平台抓取数据。随机挑选部分微博，爬取这些微博的转发用户和能够看到这些微博的非转发用户，进行用户去重和垃圾用户过滤，得到有效用户集合 U，然后爬取用户集合 U 中每个用户最近发布的100条微博，包括原创微博和转发微博，用于用户兴趣特征提取。对于获取的数据集，根据实际转发情况进行人工标注，得到最终的实验数据集，共包含286 050条数据。

微博文本具有随意、不规范的特点，因此有必要在实验前对其做如下预处理：①去噪处理：利用正则表达式去除用户历史微博文本中的超链接、hashtag 标签、@用户、标点符号等。②分词：使用汉语分词系统 Ansj 对微博文本进行分词，同时根据维基概念添加自定义词典，从而得到更加准确的分词结果。③去停用词：根据停用词集合(哈尔滨工业大学停用词表、四川大学机

器智能实验室停用词库和百度停用词表)[25],整理去重后得到停用词表,用来对微博文本进行停用词过滤。

4.2 实验设置

为了验证本文所提模型的有效性,特设置5组实验进行比较,实验具体说明如下:

方法1:使用K-means算法对用户历史微博进行聚类,得到用户兴趣主题及主题对用户的影响力;根据3.3.1-3.3.4中介绍的方法提取用户行为特征、用户交互特征、微博特征和用户兴趣特征;最后使用SVM实现微博转发预测。

方法2:方法1+情感差异特征。

方法3:文献[14]的方法。使用LDA从用户历史微博文本中抽取主题特征,结合用户特征、微博特征等其他特征运用SVM实现微博转发预测。

方法4:方法3+情感差异特征。

方法5:根据3.1中的方法构建维基知识库,对经过预处理后的用户历史微博和待预测微博按照3.2中的方法进行语义扩展,然后使用K-means算法对语义扩展后的用户历史微博进行聚类,得到用户兴趣主题及主题对用户影响力;根据3.3.1-3.3.4中介绍的方法提取用户行为特征、用户交互特征、微博特征和用户兴趣特征;最后使用SVM实现微博转发预测。

方法6(本文方法):方法5+情感差异特征。

上述6种实验方法中,方法1是为了验证语义扩展的作用自己设置的对比实验方法,与方法5的区别是:方法1是直接对用户历史微博文本进行聚类提取用户兴趣,并未进行语义扩展;方法5是对用户微博文本进行语义扩展后再进行聚类提取用户兴趣。方法1与方法5进行对比,旨在验证语义扩展对微博转发预测的有效性。方法3是将文献[14]的方法用本文的实验数据进行实验;方法2、4、6分别是在方法1、3、5的基础上加入情感差异特征,旨在探索情感差异特征对微博转发预测效果的影响;方法1与方法6进行对比,旨在探索情感差异和语义扩展共同对微博转发预测效果的影响。对6种方法的实验结果进行综合对比,全面探索情感差异特征和语义扩展对微博转发预测的作用。

4.3 实验结果及分析

本文采用十折交叉验证进行模型效果检验,即将数据集分成10等份,轮流使用其中的9份作为训练数据,剩下的1份作为测试数据,进行交叉验证,并对实验的结果取平均数,作为模型效果的估计。采用准确率P(precision)、召回率R(recall)和F值(f-measure)作为检验模型效果的评价指标。准确率用于检验模型的准确性,召回率用于检验模型的完备性,准确率和召回率相互制约,因此用F值作为模型效果的综合评价指标。其计算公式分别如下所示:

$$P=\frac{\text{被正确预测为转发的微博数}}{\text{被预测为转发的微博数}} \quad \text{公式(10)}$$

$$R=\frac{\text{被正确预测为转发的微博数}}{\text{实际被转发的微博数}} \quad \text{公式(11)}$$

$$F\text{值}=\frac{2P*R}{P+R} \quad \text{公式(12)}$$

上述6种方法的实验结果见表7和图2。通过分析实验结果可以发现,方法5在准确率、召回率上和F值上都要远远高于方法1和方法3。原因分析如下:微博属于短文本,所含特征少,存在特征稀疏问题,直接对其进行聚类效果并不理想;与传统聚类方法相比,LDA通过对文本的主题信息进行建模将文本语义特征进行有效的降维,但是LDA并不适合于微博等短文本的主题提取;方法5对微博文本进行语义扩展,挖掘潜在语义信息,解决微博文本语义稀疏问题,提高微博文本语义描述能力,然后对语义扩展后的微博文本进行聚类,能够更为准确地提取用户兴趣主题,从而提高微博转发预测的效果。

通过方法2、4、6和方法1、3、5的结果对比可以看出,加入情感差异特征后,准确率、召回率和F值均有一定程度的提升。这是因为微博内容的情感与其被转发情况存在相关性,情感差异程度大的微博往往携带更多的衍生信息,更容易得到用户的转发[2]。验证了情感差异特征对微博转发预测的有效性。

表7和图2的实验结果显示,上述6种实验方法中,本文方法(方法6)效果最好,验证了本文所提模型的有效性。这是因为用户兴趣[11-14]和微博内容的情感信息[2,15-17]对用户转发行为具有重要影响,本文提出的方法利用维基知识库对微博文本进行语义扩展,挖掘潜在语义信息,解决语义稀疏问题,然后对扩展后的用户历史微博文本进行聚类,从而更为准确地提取用户兴趣主题;同时对待预测微博进行情感分析,提取情感差异特征,考虑了微博内容的情感差异对用户转发的影响,因此效果最好。

表7 不同方法实验结果数据

比率	方法1	方法2	方法3	方法4	方法5	方法6
准确率P(%)	75.1	76.8	77.8	79.5	81.4	83.2
召回率(%)	82.6	83.8	84.9	86.2	87.3	88.8
F值(%)	78.7	80.1	81.1	82.7	84.2	85.9

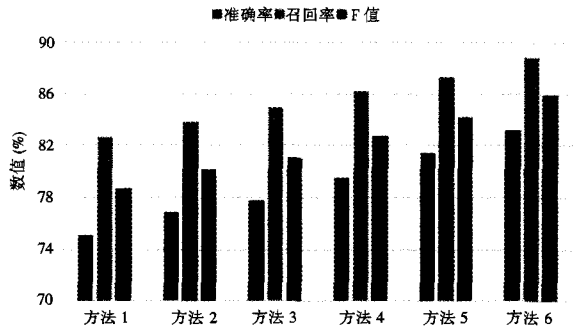

图2 不同方法实验结果比较

5 总结与展望

本文提出一个融入情感差异和用户兴趣的微博转发预测模型，利用维基知识库对微博文本进行语义扩展，解决用户兴趣表示不够全面准确的问题，同时考虑情感差异对用户转发行为的影响，在新浪微博真是数据集上进行实验，预测用户是否会对指定的某些微博产生转发行为。实验结果表明，本文方法有效提高了的微博转发预测的效果。

在未来研究中，将从以下几个方面进行改进：①结合用户标签和时间衰减动态提取用户兴趣；②研究用户活跃时间对微博转发的影响；③进一步研究用户的转发动机，探索其他特征对用户转发行为的影响，从而提高转发预测准确性。

参考文献：

[1] 张亚明，唐朝生，李伟钢. 微博机制和转发预测研究[J]. 情报学报，2013，32(8):868-876.

[2] PFITZNER R, GARAS A, SCHWEITZER F. Emotional divergence influences information spreading in Twitter[C]// International AAAI conference on weblogs and social media. Menlo Park: AAAI Press,2012:2-5.

[3] BOYD D, GOLDER S, LOTAN G. Tweet, tweet, retweet: conversational aspects of retweeting on Twitter[C].//Hawaii international conference on system sciences. Kauai:IEEE,2010:1-10.

[4] SUH B, HONG L, PIROLLI P, et al. Want to be retweeted? Large scale analytics on factors impacting retweet in Twitter network [C]// IEEE second international conference on social computing. Minneapolis:IEEE, 2010:177-184.

[5] ZHAO X, ZHU F, QIAN W, et al. Impact of multimedia in Sina Weibo: popularity and life span[C]//Joint conference of 6th Chinese semantic web symposium and the first Chinese web science conference(CSWS & CWSC). Resewch collection school of information system. New York:Springer,2013:55-65.

[6] BIAN J, YANG Y, CHUA TS. Predicting trending messages and diffusion participants in microblogging network[C]//Proceedings of the 37th international ACM SIGIR conference on research & development in information retrieval. New York: ACM Press, 2014: 537-546.

[7] 张玢玢，李兵，李岳欣. 基于特征选择的企业微博转发机制研究[J]. 情报杂志，2014(12):127-132.

[8] 张旸，路荣，杨青. 微博客中转发行为的预测研究[J]. 中文信息学报，2012，26(4):109-114.

[9] 曹玖新，吴江林，石伟，等. 新浪微博网信息传播分析与预测[J]. 计算机学报，2014(4):779-790.

[10] 刘玮，贺敏，王丽宏，等. 基于用户行为特征的微博转发预测研究[J]. 计算机学报，2016，39(10):1992-2006.

[11] 陈江，刘玮，巢文涵，等. 融合热点话题的微博转发预测研究[J]. 中文信息学报，2015，29(6):150-158.

[12] 谢婧，刘功申，苏波，等. 社交网络中的用户转发行为预测[J]. 上海交通大学学报，2013，47(4):584-588.

[13] 吴凯，季新生，刘彩霞. 基于行为预测的微博网络信息传播建模[J]. 计算机应用研究，2013，30(6):1809-1812.

[14] 李志清. 基于LDA主题特征的微博转发预测[J]. 情报杂志，2015(9):158-162.

[15] STIEGLITZ S, DANG-XUAN L. Emotions and information diffusion in social media-sentiment of microblogs and sharing behavior [J]. Journal of management information systems, 2014, 29(4): 217-248.

[16] KANAVOS A, PERIKOS I, VIKATOS P, et al. Modeling retweet diffusion using emotional content[M]. Artificial intelligence applications and innovations. Berlin:Springer,2014:101-110.

[17] NAVEED N, GOTTRON T, KUNEGIS J, et al. Bad news travel fast: a content-based analysis of interestingness on Twitter[C].// Proceeding of 3rd ACM WebSci conference. Germany: Web Science,2011:1-7.

[18] Index of /zhwiki/latest/[EB/OL]. [2016-10-20]. https://dumps.wikimedia.org/zhwiki/latest/.

[19] DKPro JWPL[EB/OL]. [2016-11-12]. https://dkpro.github.io/dkpro-jwpl/.

[20] GABRILOVICH E, MARKOVITCH S. Wikipedia-based semantic interpretation for natural language processing[J]. Journal of artificial intelligence research, 2014, 34(4):443-498.

[21] HUANG J, CHENG X Q, SHEN H W, et al. Exploring social influence via posterior effect of word-of-mouth recommendations [C]// International conference on web search and web data mining. New York:ACM Press, 2012:573-582.

[22] 徐琳宏，林鸿飞，潘宇，等. 情感词汇本体的构造[J]. 情报学报，2008，27(2):180-185.

[23] CHANG C C, LIN J. LIBSVM -- a library for support vector machine[EB/OL]. [2016-10-24]. s. http://www.csie.ntu.edu.tw/~cjlin/libsvm/.

[24] 八爪鱼采集器[EB/OL]. [2016-09-20]. http://www.bazhuayu.com/.

[25] 停用词集合(哈工大停用词表、四川大学机器智能实验室停用词库、百度停用词表)[EB/OL]. [2016-10-12]. http://

www. datatang. com/data/19300.

罗颖利：文章内容的组织和撰写。

作者贡献说明：

唐晓波：文章总体思路、架构设计；

Integrating Emotional Divergence and User Interests into the Prediction of Microblog Retweeting

Tang Xiaobo[1,2] Luo Yingli[1]

[1] School of Information Management, Wuhan University, Wuhan 430072

[2] Center for Studies of Information System, Wuhan University, Wuhan 430072

Abstract: [**Purpose/significance**] Microblog retweeting is the key way for information diffusion. The study of user retweeting behavior can better understand the information diffusion mechanism in microblog, which is of great significance for hot topic detection, public opinion monitoring and microblog marketing and so on. In this paper, we analyze the factors that affect users' retweeting behavior. The users' interest isn't comprehensive and accurate, and the influence of emotional divergence on users' is not considered in the previous research. Thus, we propose a microblog retweeting prediction model integrating emotional divergence and user interests. [**Method/process**] Firstly, we built a Wikipedia knowledge base by extracting semantic relations from Wikipedia, and extended the semantic feature of microblog text vector by using the Wikipedia knowledge base, which could solve the semantic sparse problem. To extract users' interest themes and their influence on users, we clustered the users' microblogs which was extended by Wikipedia knowledge base. Secondly, we calculated the emotional intensity of all kinds of emotions in microblog, extracting the emotional divergence features. Finally, combining users' behavior features, users' interaction features, microblog features, users' interests features and emotional divergence features, we used SVM to achieve the prediction of microblog retweeting. [**Result/conclusion**] The experimental results show that the proposed method can effectively improve the performance of microblog retweeting prediction.

Keywords: microblog retweeting user interests semantic extension emotional divergence feature sentiment analysis