基于迁移学习的唐诗宋词情感分析

吴 斌,吉 佳,孟 琳,石 川,赵惠东,李仪清
(北京邮电大学智能通信软件与多媒体北京市重点实验室,北京 100876)

摘 要: 随着计算社会学的兴起,利用数据挖掘分析社会情感是近期的研究重点. 当前的研究主要针对现代文本,对于古代诗歌这类短文本的情感分析相对较少. 本文提出了一个基于短文本特征扩展的迁移学习模型 CATL-PCO,通过分析诗歌情感对当时社会及文化进行进一步了解. 该模型首先基于频繁词对对古文特征向量进行扩展,再通过迁移学习方式,建立三个分类器并投票得出最后的情感分析结果. CATL-PCO 模型首先能够解决古文短文本特征稀疏的问题,在此基础上进一步解决由于现代译文信息匮乏所导致的古代诗歌情感分析困难问题,从而准确的分析古诗词情感倾向,从计算社会学的角度,增进对中国历史的认识. 实验表明,当训练集为中国唐诗时,本文提出方法能够准确的对唐代诗歌进行情感分类,并能应用于唐代和宋代各个时期情感分析及代表流派分析.

关键词: 情感分析; 社会计算学; 唐诗宋词; 迁移学习
中图分类号: TP393 **文献标识码:** A **文章编号:** 0372-2112 (2016)11-2780-08
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2016.11.030

Transfer Learning Based Sentiment Analysis for Poetry of the Tang Dynasty and Song Dynasty

WU Bin, JI Jia, MENG Lin, SHI Chuan, ZHAO Hui-dong, LI Yi-qing
(*Beijing Key Laboratory of Intelligent Telecommunications Software and Multimedia, Beijing University of Posts and Telecommunications, Beijing* 100876, *China*)

Abstract: With the rise of computational social science, analyzing social sentiment with data mining methods has attracted widespread attention and has become a hot spot in recent years. Existing researches of sentiment analysis mainly focus on modern text, but hardly involve the ancient short text literature. This paper proposes a short text feature extension based transfer learning model CATL-PCO(Correlation Analysis Transfer Learning-Probability Co-occurrence). Through sentiments analysis in ancient literature, this paper can discovery social and cultural development in the ancient era. CATL-PCO expands the ancient literature feature vector based on the frequent word pairs, and utilizes transfer learning method to train three sentiment classifiers. CATL-PCO solves the problem of sparsity of short text feature vector, and the scarcity of modern translation, which improves the cognition of Chinese History. Experiments demonstrate the effectiveness of the proposed method on the dataset of Chinese poems in Tang Dynasty. Moreover, different periods of Tang and Song Dynasty, and different genres are analyzed in this paper in details.

Key words: sentiment analysis; computational social science; poetries of the Tang dynasty and Song dynasty; transfer learning

1 引言

计算社会科学是社会学的分支,它利用计算机模拟、人工智能及复杂的统计方法来构建社会交互的理论模型. 目前利用计算的方法挖掘语言、词语、文字的特性已成为热门研究主题. 以哈佛大学 David Lazer 为首的 15 名知名大学教授于 2009 年 2 月在 Science(科学)杂志上发表了题为 Computational Social Science[1] 的文章提出:随着人们能够收集和分析大规模的人类行为数据并从中发现个人和群体行为的模式,一个新兴的研究领域"计算社会科学"涌现出来. 特别是近年来,随着 SNS、微博等社会化媒体的兴起,文字已经成为人们

收稿日期:2015-02-13;修回日期:2015-07-01;责任编辑:蓝红杰
基金项目:国家 973 重点基础研究发展计划(No. 2013CB329606);国家自然科学基金(No. 71231002, No. 61375058)

表达情感的主要方式. 因此通过分析文本中包含的情感,可以对人们的思想状况进行衡量,进而反映社会整体情感.

中国古代文学作品可以看做用来表达古人某一阶段思想感情的“微博”. 文献[2~4]研究了现代人们习惯使用的微博、微信等社会化媒体内容. 其中,文献[4]主要研究了国内微博这类短文本的情感倾向. 针对古代篇幅较短文学作品的研究较少,Google[5]通过研究18世纪以来的部分出版书籍,分析了其中关键词随时间的变化及其反映出的文化的变化走势. 古代文学作品的短文本特性对于其情感分析造成了一定困难. 更重要的是,虽然对于单首古诗人们可以容易理解,但批量总体性的古诗集的情感理解有必要引入机器智能.

古诗作为一种包含诸多语义且言辞简练的短文本,其情感分析存在两个挑战. 第一,中国古代诗歌本身字数很少,语言精练,如五言绝句的字数为20,其情感特征并不明显,尤其是经过数据预处理之后,得到的特征向量会更少. 第二,古文情感倾向的标注困难,人工标注的准确率不高. 对诗歌的现代译文进行情感倾向标注较为简单,更符合现代人认知. 传统机器学习要求训练集和测试集符合独立同分布假设,而迁移学习方法使用其他相近领域的知识弱化了这种假设. 正适合处理古诗词领域数据较少而现代译文领域有大量标注数据这一情况.

本文提出一个基于短文本特征扩展的迁移学习模型 CATL-PCO (Correlation Analysis Transfer Learning-Probability Co-occurrence)解决中国古代文学作品的情感分类问题. 本文首先通过关联挖掘扩充短文本特征向量,解决古代诗歌这类短文本特征稀疏问题;然后以迁移学习为主导思想,将现代译文的知识运用到无译文的古代诗歌作品中,建立两类古代诗歌的特征向量矩阵,解决由于现代译文信息匮乏所导致的古代诗歌情感分析困难问题。

2　相关工作

情感分析的目的是对带有主观性情感的文字进行分析和挖掘,其重点是情感分类. 目前的文本情感分析的方法主要为基于词典匹配的方法[6,7]和基于机器学习的方法[8]. 基于词典匹配的方法的核心是情感词典. 很多研究者已经建立了多种语言、多种情感分类的情感词典[6],同时提出了多种利用种子词库扩充情感词典的方法. 例如,文献[7]提出一种依赖扩张模型来得到情感词典的方法. 很多研究人员针对短文本也进行了大量研究,Mihalcea[12]等提出了基于语料库和知识库测量短文本片段相似性的方法. Phan[14,15]等提出将传统知识基于不同的主题转化,用于提升短文本的描述. 文献[16]进一步探索利用额外的更小的未标记的文本信息库,利用他们将短文本文件扩充为一个新的替代,这些更小的信息库不需要和已有的短文本文件符合相同的分布,长度和结构都不受限制. 上述方法对于语料库和标注较少的古代诗歌也不适用.

迁移学习被提出后受到了广泛关注,其目的是将从源领域 S 中获取的知识应用到另外一个不同却相关的目标领域 T 中去. 领域是由特征空间 X 和特征的边缘概率分布 $P(X)$ 组成的. 源领域和目标领域的特征空间和特征的边缘概率分布一般不同,或者其中某一项不一样. 迁移学习根据源领域和目标领域是否需要标注数据以及任务是否相同可以分成三类:归纳迁移学习、直推式迁移学习和无监督迁移学习[9]. 根据采用技术不同可以把迁移学习分成三类:基于权重的迁移学习、基于特征选择的迁移学习和基于特征映射的迁移学习. 基于特征映射的迁移学习的核心是特征映射. Pan等[10]通过最小化隐性语音空间上的最大均值误差来求解降维后的特征空间,随后用监督学习算法对目标领域数据进行预测. 文献[11]研究了多个相关聚类任务的学习问题,建立了一种寻找共享特征子空间的框架. 现有方法都是将源领域和目标领域映射到了新的特征空间,但在映射时会损失部分信息,对于信息匮乏的情况并不适用. 本文采用将目标特征空间映射到源特征空间中的方法,以此减少信息损失.

3　CATL-PCO 情感分析方法

3.1　问题定义

针对古诗词的特征提取,CATL-PCO 模型采用传统 TF-IDF 方法计算文本特征. 给定大量目标领域古诗词短文本数据 T 和源领域 S,通过基于关联分析的迁移学习方式形成的分类器 $F_{final}(.)$ 将 T 中古诗进行情感分类,得出分类结果. 具体过程可以表示成以下形式:

$$Classsify(T) = F_{final}(T,S) \tag{1}$$

将大量古代诗歌数据作为目标领域数据 $T = \{\boldsymbol{t}_l, c_l\}_{l=1}^{L}$,其中 $\boldsymbol{t}_l \in R^{M_T}$ 表示从第 l 篇古文中提取的特征向量,$C_l \in R^C$ 表示该诗词的对应类别,$R^C = \{$正,中,负$\}$. T 中共有 L 首古诗,其分类未知. 与部分诗词相对应的现代译文作为源领域的数据 $S = \{\boldsymbol{s}_k, C_k\}_{k=1}^{K}$,其中 $\boldsymbol{s}_k \in R^{M_S}$ 表示从第 k 篇译文中提取的现代文特征向量,C_k 表示该诗词的对应类别,源领域 S 中共有 K 篇现代译文,K 远小于 L. S 中译文分类已知,即对应的 K 篇古代诗歌的分类已知,分别表示为 $T_{label} = \{\boldsymbol{t}_k, c_k\}_{k=1}^{K}$ 和 $T_{left} = \{\boldsymbol{t}_l, c_l\}_{l=1}^{L-K}$. 源领域和目标领域的特征向量分别为 $\boldsymbol{s}$ 和 $\boldsymbol{t}$,向量大小分别是 M_S 和 M_T.

3.2　情感分类模型 CATL-PCO

CATL-PCO 模型首先采用基于 FP-Growth 的关联分

析方法将古诗本身特征向量 $\boldsymbol{t}$ 扩充为 $\boldsymbol{t}'$,随后采用基于特征映射的迁移学习方法,提出两种特征映射算法,将特征从目标特征空间 T 映射到源特征空间 S 中,并利用其知识. 通过数据 T 和 S,建立特征共现矩阵 $\boldsymbol{CO} \in R^{M_S * M_T}$,然后计算出带权重的条件概率矩阵 $\boldsymbol{PCO}$. 通过 $\boldsymbol{PCO}$ 将 $\boldsymbol{t}'$ 映射到源特征空间 S 中. 映射后得到的特征分别定义为是 $\boldsymbol{S}_{E\text{-}PCO}$ 和 $\boldsymbol{S}_{P\text{-}PCO}$. 借鉴 boost 算法的思想,通过 $\boldsymbol{t}'$、$\boldsymbol{S}_{E\text{-}PCO}$ 和 $\boldsymbol{S}_{P\text{-}PCO}$ 分别训练 $F_t(.)$、$F_{SE}(.)$ 和 $F_{SP}(.)$ 三个分类器. 通过三个分类器投票表决,得到最终分类器 $F_{final}(.)$ 的结果,下面是方法的描述.

算法 1 基于迁移学习的中国古代诗歌情感分析方法

输入:T_{label},S

输出:$F_{final}(.)$

1:计算特征共现矩阵 $\boldsymbol{CO}$

2:将古诗本身特征向量 $\boldsymbol{t}$ 通过 *FP-Growth* 扩充成 $\boldsymbol{t}'$

3:计算带权重的条件概率矩阵 $\boldsymbol{PCO}$

4:构建新的特征表达 $\boldsymbol{S}_{E\text{-}PCO}$ 和 $\boldsymbol{S}_{P\text{-}PCO}$

5:利用 $\boldsymbol{t}$、$\boldsymbol{S}_{E\text{-}PCO}$ 和 $\boldsymbol{S}_{P\text{-}PCO}$ 三种特征训练 3 个分类器 $F_t(.)$、$F_{SE}(.)$ 和 $F_{SP}(.)$

6:三种分类器投票表决,获得总分类器 $F_{final}(.)$ 的结果

3.3 基于关联分析模型的特征扩充

由于古代诗歌这类短文本的特征向量较少,描述情感特征较弱,使得在对诗歌进行情感分类的过程中,其具体情感不能很明显的表示出来. 本文采用 FP-Growth 方法来对古文进行特征向量的扩展. FP-Growth 算法核心思想是将数据按照一定规则压缩到频繁模式树中,然后在树中求解频繁项目集合. 它是关联规则经典算法之一.

以源领域和目标领域中所有古诗为背景,采用 FP-Growth 算法进行关联挖掘. 诗集 $P=\{p_1,p_2,\cdots,p_n\}$ 中,每篇古诗的词集为 $T=\{t_1,t_2,\cdots,t_n\}$,$C=\{$正,中,负$\}$ 为每首诗所对应的类别. 为了避免扩展的特征向量中出现对古诗情感分类的干扰词,本文对于频繁词对的关联分析是建立在不同情感分类的古诗的基础上来进行的. 具体步骤如下:

(1)计算频繁词对的全局支持度 $S(P)$:$S(P)=\dfrac{N(P_w)}{N(P)}$,其中 $N(P_w)$ 表示诗集 P 中包含词对 w 的诗的个数,$N(P)$ 表示诗集中古诗的个数.

(2)计算频繁词对的置信度 $C(W,P)$:$C(W,P)=\dfrac{N(P_{(w,p)})}{N(P_w)}$,其中 $N(P_w)$ 表示诗集 P 中包含词对 w 的诗的个数,$N(P_{(w,p)})$ 表示包含词对 w 且类别为 c 的诗的个数.

(3)将古诗 p_1 的特征向量$(\boldsymbol{t}_{p_1},\boldsymbol{t}_{p2},\cdots,\boldsymbol{t}_{pn})$针对得到的频繁词对进行特征向量匹配,得到新增特征向量$(\boldsymbol{t}'_{p_1},\boldsymbol{t}'_{p2},\cdots,\boldsymbol{t}'_{pn})$,将匹配到的特征向量添加到原特征向量后,形成扩充后的特征向量$(\boldsymbol{t}_{p_1},\boldsymbol{t}_{p2},\cdots,\boldsymbol{t}_{pn},\boldsymbol{t}'_{p_1},\boldsymbol{t}'_{p2},\cdots,\boldsymbol{t}'_{pn})$.

通过设定支持度阈 a,可以得到全局支持度大于 a 的频繁词对;通过设置置信度阈 b,可以得到特征词对 w 属于类别 c 的可能性大小. 本文采用 FP-Growth 算法针对有相同情感分类的二元词对进行关联挖掘,得到多个二元频繁词对. 之后对古诗进行特征向量扩充,进而扩展这类短文本的特征向量.

3.4 特征共现矩阵

本文采用 TF-IDF 方法来提取特征,计算 TF-IDF 时采用该词在每类诗文中的逆文档频率,得到正、中、负三类诗文中出现的词的 TF-IDF 排序,选取各类排名在前 20% 的词,按排序组成词典. 词典中词的出现的次数作为特征 $\boldsymbol{s}$ 和 $\boldsymbol{t}$. 从古文中提取出的词典 D_T 共有 M_T 个词,即 $\boldsymbol{t}$ 由 M_T 个非负整数组成;从现代译文中提取出的词典 D_S 共有 M_S 个词,即 $\boldsymbol{s}$ 由 M_S 个非负整数组成. 根据 $\boldsymbol{s}$ 和 $\boldsymbol{t}$ 建立共现矩阵 $\boldsymbol{CO} \in R^{M_S * M_T}$ 来表示古文和现代文的相关性. $\boldsymbol{CO}$ 以 D_S 和 D_T 中的词作为矩阵的边,大小为 $M_S * M_T$,矩阵中的值 e_{ij} 为目标领域中词 t_i 与原领域中词 s_j 出现在同一首诗文中的次数.

$$e_{ij} = \sum_{k=1}^{K} \delta_k(t_i,s_j) \tag{2}$$

其中 $\delta_k(t_i,s_j)=\begin{cases}1, & t_i \in k \cap s_j \in k\\ 0, & \text{否则}\end{cases}$,$e_{ij}$ 为共现矩阵 $\boldsymbol{CO}$ 中第 i 行 j 列的元素的值,$i \in [1,M_T]$ 表示 D_T 中第 i 个词,$j \in [1,M_S]$ 表示 D_S 中第 j 个词. 对于每首诗词 k,如果其译文特征包括 s_j 且古文特征包括 t_i,则 δ 函数值为 1,否则为 0.

3.5 条件概率矩阵

本文提出的两种映射方法 $\boldsymbol{S}_{E\text{-}PCO}$ 和 $\boldsymbol{S}_{P\text{-}PCO}$ 均是以条件概率矩阵 $\boldsymbol{PCO}$ 为基础. 对于共现矩阵 $\boldsymbol{CO}$,计算每个源特征空间中特征出现的条件概率,

$$\begin{aligned} P(s_j \mid t_i) &= \frac{P(t_i,s_j)}{P(t_i)} \\ &= \frac{C(t_i,s_j)}{\sum_{j}^{M_S} C(t_i,s_j)} = \frac{e_{ij}}{\sum_{j}^{M_S} e_{ij}} \end{aligned} \tag{3}$$

其中 $C(t_i,s_j)$ 表示译文特征 s_j 和古文特征 t_i 共同出现次数,即 e_{ij}. 由于 e_{ij} 数值较小,本节引进权重 W_{ij} 来提高在两个词典中排名较高的词的共现关系的重要程度. 基于权重的 $C(t_i,s_j)$ 如下:

$$C_W(t_i,s_j) = C(t_i,s_j) * w_{ij} = C(t_i,s_j) * \frac{1}{\left(\frac{i*\alpha}{M_T}+1\right)\left(\frac{j*\alpha}{M_S}+1\right)} \quad (4)$$

其中,i 为 D_T 中的第 i 个词,M_T 为 D_T 的总词数,α 表示将 D_T 中的词分为 α 个等级,为避免分母为 0 的情况,在每个子式中 +1. $C_W(t_i,s_j)$ 的意义是将词典中的词分为 α 个等级,排在前 M_T/α 的词的权重为 1,接下来 M_T/α 个词的权重为 1/2,直到最后的 M_T/α 个词的权重为 $1/\alpha$.

则 $P(s_j|t_i)$ 转化为 $P_W(s_j|t_i)$,即

$$P_W(s_j|t_i) = \frac{C_w(t_i,s_j)}{\sum_j^{M_S} C_w(t_i,s_j)} = \frac{e_{ij}*w_{ij}}{\sum_j^{M_S} e_{ij}*w_{ij}} \quad (5)$$

由 $P_W(s_j|t_i)$ 组成概率矩阵 $\boldsymbol{PCO} \in R^{M_T*M_S}$,矩阵 $\boldsymbol{PCO}$ 和矩阵 $\boldsymbol{CO}$ 规模一样,均是 M_S*M_T,不同的是元素由 e_{ij} 变成了 $P_W(s_j|t_i)$.

3.6　构建新的特征表达

对于古诗词来讲,其扩展后的古文特征 $\boldsymbol{t}'$ 是可以获得的,但其现代译文特征 $\boldsymbol{s}$ 是不确定的. 很多古诗由于缺少译文而无法获得 $\boldsymbol{s}$. 针对这一情况,本文提出两种迁移方法:基于期望的条件概率矩阵和基于概率的条件概率矩阵,以此通过古文特征与之前获得的条件概率矩阵获得其在现代译文特征空间中的映射.

基于期望的条件概率,通过扩充后的古文特征 $\boldsymbol{t}'$ 映射为现代译文特征的期望,即根据扩充后的古文特征 $\boldsymbol{t}'$ 来映射 $\boldsymbol{s}$ 中每个特征的值,这些特征组成 $\boldsymbol{S}_{E\text{-}PCO}$,长度为 M_S,$\boldsymbol{S}_{E\text{-}PCO}$ 中的值为预测的词频.

$$\boldsymbol{S}_{E\text{-}PCO} = E'(s_j) = Q' * P'(s_j) = Q' * \sum_i^{M_T} P'(t_i)P'(s_j|t'_i) = \sum_i^{M_T} C'(t'_i)P'(s_j|t'_i) \approx \sum_i^{M_T} C'(t_i)P'_W(s_j|t'_i) \quad (6)$$

其中,$E'(s_j)$ 表示新的古诗对应的现代特征中词 s_j 的期望,Q' 表示新的古诗特征总和,$P'(t_i)$ 表示新的古诗特征中词 t'_i 的比例,$C'(t'_i)$ 表示新的古诗特征中词 t_i 出现的词频,即 $\boldsymbol{t}'$ 中的值. 对于等式中的 $P'(s_j|t'_i)$,本文用先前的知识 $P_W(s_j|t'_i)$ 来近似. 这样就将古文特征空间中的数据迁移到现代文特征空间中,接下来可以利用现代译文的知识来分类.

基于概率的条件概率,是通过古文特征 $\boldsymbol{t}$ 映射为现代译文特征的概率,即根据古文特征 $\boldsymbol{t}'$ 来映射 s_j 中每个特征可能出现的概率,这些特征组成 $\boldsymbol{S}_{P\text{-}PCO}$,长度为 M_S,$\boldsymbol{S}_{P\text{-}PCO}$ 里的值即为预测的出现可能性,范围是[0,1]. 这里 $\boldsymbol{t}'$ 的值需要修改为 0 或 1,即词 t_i 出现为 1,否则为 0.

$$\boldsymbol{S}_{P\text{-}PCO} = P'_C(s_j) = \sum_i^{M_T} P'_C(t'_i) * P'(s_j|t'_i) \approx \sum_i^{M_T} P'_C(t'_i) * P'_W(s_j|t'_i) \quad (7)$$

其中,$P'_C(s_j)$ 表示新的古诗对应的现代特征中词 s_j 出现的概率,$P'_C(t'_i)$ 表示新的古诗特征中词 t'_i 的出现概率,即 0 或 1,$P'(s_j|t'_i)$ 由 $P'_W(s_j|t'_i)$ 来近似.

3.7　分权表决

通过上面的方法一首古诗可以得到 $\boldsymbol{t}'$、$\boldsymbol{S}_{E\text{-}PCO}$ 和 $\boldsymbol{S}_{P\text{-}PCO}$ 三种特征,可以训练出 3 个分类器 $F_t(.)$、$F_{SE}(.)$ 和 $F_{SP}(.)$. 本文采用人工神经网络中的 RBF Network 分类器. RBF Network 是一种采用径向基函数(Radial basis function,RBF)的人工神经网络方法. 该方法具备较强的输入和输出映射功能、其学习过程收敛速度快、分类能力强、网络连接权值与输出呈线性关系等优点,所以本文采用 RBF Network 作为分类器,其中径向基核函数 φ 取高斯函数. 分类器基本函数为

$$F_t(.) = \sum_{i=1}^{h} \omega_{ij}\varphi(\|x_t - c_i\|) \quad (8)$$

$$F_{SE}(.) = \sum_{i=1}^{h} \omega_{ij}\varphi(\|x_{S_{E\text{-}PCO}} - c_i\|) \quad (9)$$

$$F_{SP}(.) = \sum_{i=1}^{h} \omega_{ij}\varphi(\|x_{S_{P\text{-}PCO}} - c_i\|) \quad (10)$$

根据三种分类器的投票表决结果来确定该首诗歌的情感分类,即通过 $F_t(.)$、$F_{SE}(.)$ 和 $F_{SP}(.)$ 的结果来确定总分类器 $F_{final}(.)$ 的结果. 表决公式如下,

$$F_{final}(.) = \begin{cases} C_k, & F_t(.) = F_{SE}(.) = F_{SP}(.) = C_k \\ C_k, & \text{两个 } F(.) = C_k \cap \text{另一个 } F(.) \neq C_k \\ F_t(.), & \text{三个 } F(.) \text{均不相等} \end{cases} \quad (11)$$

即少数服从多少,如果三个结果都不同,取 $F_t(.)$ 的结果.

4　诗词情感分析方法实验及分析

4.1　数据来源

本节从互联网中的文学网站获取大量唐代诗词数据,得到数据集. 具有来源为八斗文学(http://poem.8dou.net/)和古诗文网(http://www.gushiwen.org/),收集到大量诗词数据,共计 253197 首,其中唐朝 45497 首,宋朝 211700 首. 邀请三名研究人员对其中 950 首唐诗进行了人工标注,将诗词标注为正、中、负三类情感倾向,标注结果为正面情感 386 首,中性情感 212 首,负面情感 352 首. 本实验的硬件环境为:Intel Pentium Dual T3400 2,16GHz,2G 内存. 软件环境为:Windows XP

系统.

4.2 情感分类实验

本节将基于迁移学习的情感分类 TL-PCO 方法[13]与本文提出 CATL-PCO 方法进行对比实验. 实验阶段使用 $\boldsymbol{t}$、$\boldsymbol{S}_{E\text{-}PCO}$ 和 $\boldsymbol{S}_{P\text{-}PCO}$ 三种特征建立 $F_t(.)$、$F_{SE}(.)$ 和 $F_{SP}(.)$ 三个基础分类器. 以上方法均与本文提出方法 $F_{final}(.)$ 的进行了对比实验.

4.2.1 古诗特征向量扩展分析

本节实验主要针对古诗这类短文本进行特征向量扩展,以此弥补其特征向量不足导致的情感分类不明确的问题,通过基于 FP-Growth 关联挖掘得到的频繁词对,将对古诗原特征向量进行特征向量扩展. 经过多次实验讨论频繁词对支持度和置信度这两个参数的取值. 实验首先设定置信度不同,支持度相同的参数进行实验,结果表明,支持度相同的情况下,置信度取不同的值对于实验结果影响很小. 随后,实验选取置信度为 7%,选取不同的支持度进行实验,生成不同的频繁词对,进而生成不同长度的扩展后的古诗词,并以 $F_t(.)$ 为分类器,分别进行情感分析实验,实验结果如图 1.

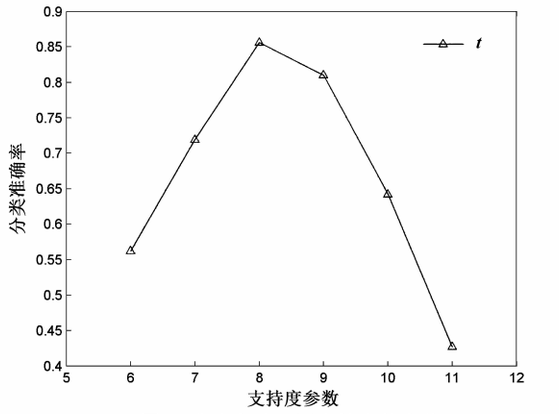


图1 准确率随不同支持度的变化

由图 1 可知,当支持度 a 为 8%,置信度为 7% 时,基于古文特征建立的特征向量的分类效果较为准确,所以实验选取支持度 a 为 8%,置信度为 7%.

使用 FP-Growth 进行频繁词对匹配后,本文将每一首古诗分词之后的结果在频繁词对库中进行匹配,从而得到扩展的古诗词特征向量. 举例来说,一首古诗在没有进行特征向量扩展之前的分词结果如下:红豆 生 南国 春来 发 几 枝 愿 君 多 采撷 此 物 最 相思. 扩展之后的结果如下:红豆 生 南国 春来 发 几 枝 愿 君 多 采撷 此 物 最 相思 独 春 无 山别 重 如 客 落 知 犹 空 愁 老 未 闻 过 来 更 得 万 一 里 人 不去 为 归 夜 中 长 言 我 入 年 已 岂 生 月 行 欲 有 风 更 上 已 是 飞 还 烟 心 声 前 时 何 明 家 雨 见 花 秋,其中,经过频繁词对匹配之后添加的词中有以下几个词语,如“别”、“落”、“愁”、“老”等,均能够表达出原来古诗词中的相思之情,加深了该诗的情感倾向.

4.2.2 情感分类实验准确率

本节实验结果均为使用不同参数进行 20 次 10 折交叉验证方式得出结果的平均值. F_t 代表通过古文扩展特征进行机器学习的方法,F_SE 代表基于期望的迁移学习方式,F_SP 代表基于概率的迁移学习方式,F_final 代表基于特征扩展和迁移学习的分权表决方法.

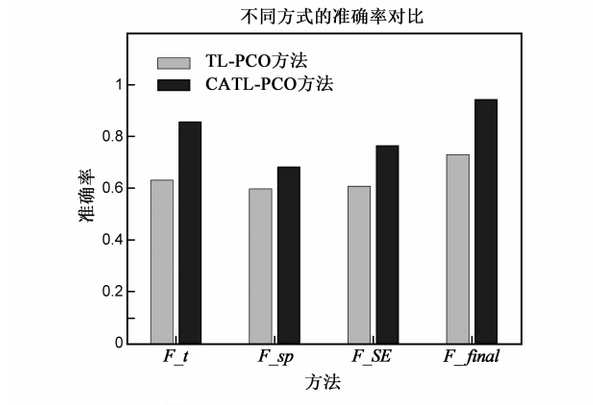


图2 准确率随不同特征的变化

由图 2 可知,当单独使用 F_t、F_SE 和 F_SP 这三种特征进行古诗情感分类时,分类准确率均较高,均在 70% 以上,其中直接使用 $\boldsymbol{t}$ 时效果最好,准确率超过 85%,这也是在表决中三者意见均不一样时以 $F_t(.)$ 为主的原因. 本文提出 CATL-PCO 方法的分权表决方式,将准确率提升至 94.3%,较单纯使用古文特征 $\boldsymbol{t}$ 进行情感分类的准确率提高了 8%. 两种方法比较可知,由于 CATL-PCO 方法着重关注古诗词的短文本特性,对古诗进行特征向量扩展后,情感分类的准确率有明显提升,可见本文提出基于关联分析的迁移学习情感分类方式的有效性.

4.2.3 条件概率矩阵的权重实验

CATL-PCO 方法中条件概率矩阵计算时需要权值参数 α,α 决定的权值对于共现矩阵中值的作用起到很大影响. 对于 α 值的确定,本节采用实验方法来验证,α 取值为[1,45],对比其准确率,准确率均为 20 次 10 折交叉验证的结果. $\alpha=1$ 表示不带有权重. 由图 3 可知,对于 $\boldsymbol{S}_{E\text{-}PCO}$ 和 $\boldsymbol{S}_{P\text{-}PCO}$ 方式,α 的取值对于准确率的影响不大,我们对 $\boldsymbol{S}_{E\text{-}PCO}$ 方式选取 $\alpha=1$,对于 $\boldsymbol{S}_{P\text{-}PCO}$ 选取 $\alpha=25$.

4.2.4 RBF Network 高斯参数

本文提出的 CATL-PCO 方法,在情感分类过程中使用 RBF Network 进行情感分类,RBF 神经网络输出层选择函数分别为 $F_t(.)$、$F_{SE}(.)$ 和 $F_{SP}(.)$,径向基核函数 φ 取高斯核函数. 分类器基本函数为

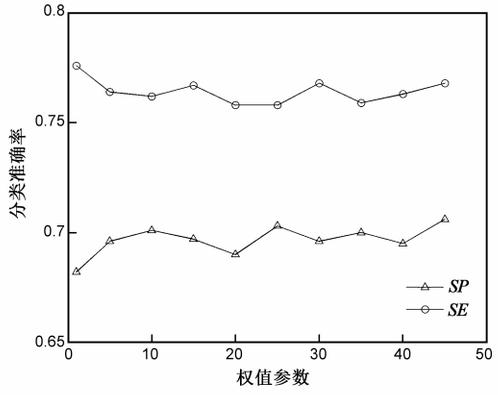


图3　准确率随α的变化

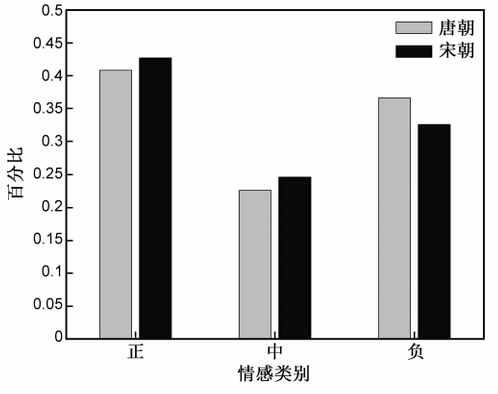


图4　唐代和宋代情感类别

$$\varphi(x,c_i)=e^{\left(\frac{-\|x-c_i\|^2}{\sigma_i^2}\right)},i=1,2,\cdots n \qquad (12)$$

实验中认为其中的高斯参数 σ 的取值对实验结果影响较小,所以设定 $\sigma=e^{\frac{1}{2}}$.

4.3　实例研究

本节实验从收集到的253197首诗歌数据中随机选取70827首诗词,其中唐朝诗歌35543首,宋朝诗歌35284首. 对这些诗词通过本文提出CATL-PCO方法进行情感分类. 实例研究的目的是利用CATL-PCO模型对海量古诗词进行情感倾向分类,从而加深对历史的了解. 实验首先使用CATL-PCO模型对大量古诗词进行特征扩展,增加其特征向量的长度,在此基础上利用迁移学习的方法,将译文映射到古文当中去,从而更加准确的判断七万余首古诗词的情感分类.

4.3.1　唐代诗词总体评价

本节首先利用CATL-PCO方法从整体上对唐代35543首诗歌和宋朝35284首诗歌进行情感分类,实验结果如图4. 由图4可知,唐朝和宋朝诗人的普遍情感较高. 其中对于唐朝诗歌的情感分类结果与文献[13]中提出的TL-PCO方法相比,CATL-PCO方法对于唐朝诗歌的正向情感预测较高,这是因为该方法针对诗歌这类短文本的特征向量扩充,解决了古代诗歌特征向量稀疏等问题,扩充了诗歌的特征向量,加强了诗歌的情感倾向.

4.3.2　唐代、宋代诗词应用分析

唐朝主要分为四个时期:初唐、盛唐、中唐和晚唐. 宋朝主要分为北宋和南宋. 实验在每个时期选取代表性诗人,以此来代表不同时期的诗歌情感状况. 实验中由于诗词的数量巨大,本节实验对选取的诗歌中具有代表性诗人作品,结合朝代、流派、地位、处境等因素,分析情感分类.

唐代各时期诗词情感变化分类结果如图5所示. 从实验数据可以看出唐代各时期诗词的情感走向,正向诗词比例除晚唐之外均较高,中性诗词所占比例变化不大,负向诗词所占比例和正向正好相反. 总体上看,实验数据中初唐、盛唐和中唐时期的幸福度总体高于与晚唐,这与历史状况相符.

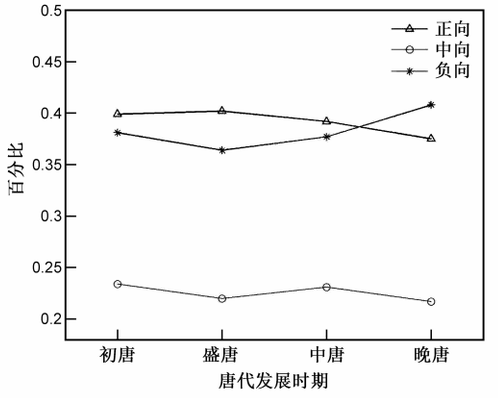


图5　唐代各时期诗词情感变化

图6表示宋代各时期诗词情感变化分析结果. 从图6来看,对于宋朝来说,南宋的正向情感较北宋略高,这是因为,南宋是中国历史上经济发达、文化繁荣、科技进步的朝代. 两个时期的负向情感基本持平. 可见以上两图中的曲线走向与历史发展一致.

图7表示盛唐两个流派情感对比结果. 唐朝的诗派主要以山水田园诗派和边塞诗派. 本实验选取两个诗派的代表诗人的作品,对两个诗派进行对比分析. 从实验数据结果图7中可以发现,山水田园派诗人的情感较为正向,负向较低. 这是因为山水田园派诗人更多的显示出宁静闲适的精神状态. 而边塞派诗人更多表现征人离妇的思想感情.

图8表示宋朝三个流派情感对比结果. 由图8可知,实验数据中宋初唐晚派诗人的情感较为正向,负向较低. 昌黎诗派正负情感相差相对较小. 实验数据中荆公诗派在早期以正向情感居多,后期以负向情感居多.

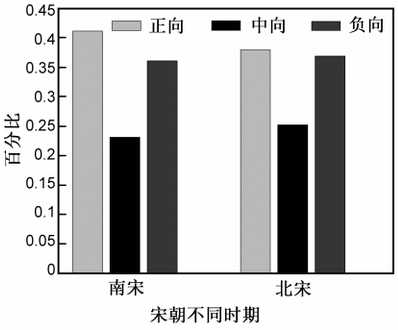


图6 宋代各时期诗词情感变化

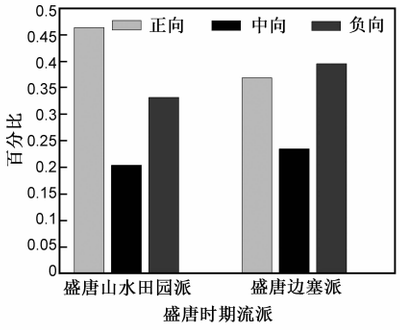


图7 盛唐两个流派情感对比

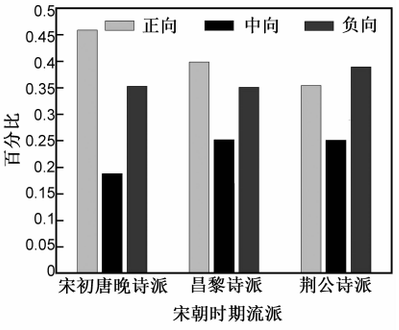


图8 宋朝三个流派情感对比

5 结束语

作为古代微博的一种表现形式,古代诗歌是研究时代概貌的良好素材. 为了分析古代诗词中表达的情感,本文提出了一种基于短文本特征扩展的迁移学习情感分析方法 CATL-PCO. 首先通过关联分析方法,弥补了古代诗歌的短文本特性稀疏对于情感分析所造成的挑战. 其次通过迁移学习中特征映射方法,合理地利用了古文和现代译文的知识,建立三个分类器,通过分权表决的方法得到最终结果. 实验表明,CATL-PCO 方法能够有效的对古代诗歌进行情感的分类及分析,对古文进行基于关联分析的特征扩展方式较传统机器学习方法的效果有很大提升,在特征向量扩展后的古文基础上进行迁移学习的情感分析方法较单纯基于迁移学习的情感分析方式在对大量数据集进行情感分析的准确率上有所提升. 随后,本文利用提出的情感分析方法,分析了中国唐诗宋词的种种方面,结合相关文学研究,证实了分析结果的合理性. 因此,本文提出的 CATL-PCO 方法能够在一定程度上对诗词等历史文献中的情感进行分析,将情感分析方法适用的范围扩展到了诗词文献的范围之中,为情感分析领域拓展了新的研究道路. 后续工作可以考虑更多的分析特征如诗人的年龄、性别、地位等,还可通过构建诗人和诗歌的异质网络等扩展分析角度和方法.

参考文献

[1] DLazer, A S Pentland, L. Adamic, S Aral, et al. Life in the network: the coming age of computational social science [J]. Science, 2009, 323(5915): 721.

[2] Nakov P, Kozareva Z, Ritter A, et al. Semeval-2013 task 2: Sentiment analysis in twitter[A]. In Proceedings of the International Workshop on Semantic Evaluation (SemEval 2013)[C]. Dublin: Association for Computational Linguistics, 2013. 312 – 320.

[3] Fu X, Liu G, Guo Y, et al. Multi-aspect sentiment analysis for Chinese online social reviews based on topic modeling and HowNet lexicon[J]. Knowledge-Based Systems, 2013, 37: 186 – 195.

[4] 刘楠. 面向微博短文本的情感分析研究[D]. 武汉大学, 2013.
LIU Nan, The research of microblogging short text oriented sentiment analysis[D]. Wuhan Univeristy, 2013. (in Chinese)

[5] Michel J B, Shen Y K, Aiden A P, et al. Quantitative analysis of culture using millions of digitized books[J]. Science, 2011, 331(6014): 176 – 182.

[6] Dong Z, Dong Q. HowNet-a hybrid language and knowledge resource [A]. Natural Language Processing and Knowledge Engineering[C]. IEEE, 2003. 820 – 824.

[7] Liang J, Tan J, Zhou X, et al. Dependency Expansion Model for Sentiment Lexicon Extraction[A], Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2013 IEEE/WIC/ACM International Joint Conferences on[C]. IEEE, 2013, 3: 62 – 65.

[8] Liu B. Sentiment analysis and opinion mining[J]. Synthesis Lectures on Human Language Technologies, 2012, 5(1): 1 – 167.

[9] 张景祥, 王士同, 邓赵红, 等. 具有协同约束的共生迁移学习算法研究[J]. 电子学报, 2014, 42(3): 556 – 560.
ZHANG Jing-xiang, WANG Shi-tong, DENG Zhao-hong, et al. Symbiosis transfer learning method with collaborative constraints[J]. Acta Electronica Sinica, 2014, 42(3): 556 – 560. (in Chinese)

[10] Pan S J, Kwok J T, Yang Q. Transferlearning via dimensionality reduction[A], Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence[C]. AAAI Press, 2008. 677 – 682.

[11] Gu Q, Zhou J. Learning the shared subspace for multi-task clustering and transductive transfer classification[A]. Data Mining, 2009. ICDM09. Ninth IEEE International Conference on[C]. IEEE, 2009. 159 – 168.

[12] R Mihalcea, C Corley, C Strapparava. Corpus-based and

knowledge-based measures of text semantic similarity [A]. In Proceedings of the 21st National Conference on Artificial Intelligence - Volume 1 [C]. AAAI Press, 2006. 775 – 780.

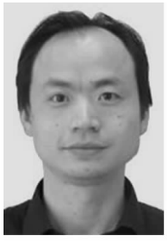
[13] Zhao Huidong, Wu Bin. Sentiment analysis based on transfer learning for Chinese ancient literature [A]. Behavior, Economic and Social Computing (BESC), 2014 International Conference on[C]. IEEE, 2014. 1 – 7.

[14] Xuan-Hieu Phan, Cam-Tu Nguyen, Dieu-Thu Le, Le-Minh Nguyen, Susumu Horiguchi, Quang-Thuy Ha. A hidden topic-based framework toward building applications with short web documents[J]. IEEE Transactions on Knowledge & Data Engineering, 2011, 23(7): 961 – 976.

[15] X-H Phan, L-M Nguyen, S. Horiguchi. Learning to classify short and sparse text & web with hidden topics from large-scale data collections[A]. In Proceeding of the 17th international conference on World WideWeb, WWW' 08 [C]. ACM, 2008. 91 – 100.

[16] Petersen H, Poon J. Enhancing short text clustering with small external repositories[A]. Proceedings of the Ninth Australasian Data Mining Conference [C]. Australian Computer Society, Inc, 2011. 79 – 90.

作者简介



吴　斌　男,1969 年生,湖南长沙人,教授、博士生导师. 2002 年中国科学院计算技术研究所博士毕业. 主要从事复杂网络、数据挖掘、海量数据并行处理、可视分析、电信客户关系管理等方面的研究工作.
E-mail: wubin@ bupt. edu. cn

吉　佳　女,1989 年生,辽宁鞍山人,北京邮电大学硕士研究生. 主要研究领域为数据挖掘与物联网大数据.

孟　琳　女,1993 年生,山东莱芜人,2015 年在北京邮电大学获学士学位,现为北京邮电大学计算机学院硕士研究生. 主要研究领域为数据挖掘.