软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2013,24(7):1545−1556 [doi: 10.3724/SP.J.1001.2013.04278]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

多维度等级评分模型优化技术*

王会珍[1,2], 朱靖波[1,2]

[1](医学影像计算教育部重点实验室(东北大学),辽宁 沈阳 110819)

[2](东北大学 自然语言处理实验室,辽宁 沈阳 110819)

通讯作者: 王会珍, E-mail: wanghuizhen@mail.neu.edu.cn

摘 要: 研究了多维度等级评分模型的训练学习优化技术.为了解决不同用户之间的评分标注所存在的不一致性,提出两种简单、有效的模型训练优化技术,包括基于容忍度的样本选择方法和基于排序损失的样本选择方法.另外,为了充分利用不同特征的用户评分标注之间的相关性,提出了一个面向属性的协同过滤技术以改善多维度等级评分模型.在两个公开的英语和汉语真实餐馆评论数据集上进行实验验证,实验结果表明,所提出的方法有效地改善了等级评分的性能.

关键词: 排序学习;有序回归模型;多维度等级评分模型;情感分析;协同过滤

中图法分类号: TP391 **文献标识码**: A

中文引用格式: 王会珍,朱靖波.多维度等级评分模型优化技术.软件学报,2013,24(7):1545−1556. http://www.jos.org.cn/1000-9825/4278.htm

英文引用格式: Wang HZ, Zhu JB. Optimizations of multi-aspect rating inference model. Ruan Jian Xue Bao/Journal of Software, 2013,24(7):1545−1556 (in Chinese). http://www.jos.org.cn/1000-9825/4278.htm

Optimizations of Multi-Aspect Rating Inference Model

WANG Hui-Zhen[1,2], ZHU Jing-Bo[1,2]

[1](Key Laboratory of Medical Image Computing, Ministry of Education (Northeastern University), Shenyang 110819, China)

[2](Natural Language Processing Laboratory, Northeastern University, Shenyang 110819, China)

Corresponding author: WANG Hui-Zhen, E-mail: wanghuizhen@mail.neu.edu.cn

Abstract: This paper addresses an issue of training optimization of multi-aspect rating inference. First, to address the issue of author inconsistency rating annotation, this paper proposes two simple approaches to improving the standard rating inference models by optimizing sample selection for training, including tolerance-based selection and ranking-loss-based selection methods. Second, to explore correlations between ratings across a set of aspects, this paper presents an aspect-oriented collaborative filtering technique to improve rating inference models. Experiments on two publicly available English and Chinese restaurant review data sets have demonstrated significant improvements over standard algorithms.

Key words: learning to rank; ordinal regression model; multi-aspect rating inference model; sentiment analysis; collaborative filtering

目前,互联网上用户评论资源规模飞速增长,如对产品的评论或个人博客.情感分析(sentiment analysis)和观点挖掘(opinion mining)逐渐成为网络信息智能处理领域的一个热点研究方向,其目的是从网上丰富的观点性文本中自动挖掘出人们对于某社会问题或者某产品的褒贬倾向性评价[1].但在很多实际应用中,如电影业,产品生产者可能更关心人们对于其产品的细颗粒度评价(如等级评分),而不仅仅是简单的褒贬性二元评价.简单来说,等级评分(rating inference)机制比褒贬评价(polarity analysis)机制更能体现具体和准确的用户观点倾向性评

* 基金项目: 国家自然科学基金(61073140, 61100089); 高等学校博士学科点专项科研基金(20100042110031); 中央高校基本科研业务费专项资金(N110404012)

收稿时间: 2011-09-29; 修改时间: 2012-02-15; 定稿时间: 2012-07-03

价,如 5 分制等级评分机制.目前,很多民意调查机构都采用了(5 或 10 分制)等级评分机制来调查用户的观点评价情况.但在实际调查过程中,很多用户不太愿意只是通过对事先给定的提问进行简单打分来准确表达自身的观点评价,而是希望采用文字性评论来准确表达自身的具体观点评价.另外,针对有些社会话题或产品的社会调查,设计合适的提问及等级评分机制的调查问卷也是不太容易的.并且,一旦所调查的评价机制和所设计的观点问题发生变化,我们需要重新开展新的调查工作、重新设计调查问卷,从而导致整个过程的人工代价非常高.由此而产生一个有意义的研究课题:如何从用户观点自由文本中自动挖掘用户观点评价满意度情况,即等级评分技术[2−4].这样的话,当每次需要开展新的调查工作时,只需要重新执行用户观点自动挖掘/分析系统就可以实现,不需要重新收集用户的评价数据和重复代价昂贵的调查过程.因此,如何利用机器学习算法来分析顾客评价观点的满意度(等级评分)是值得研究的一个问题.等级评分能够反映出用户对某些社会问题或者产品的评价(满意度).等级评分技术的目标是,从网络评价文本中自动挖掘评价者的真实观点,并以等级分数的形式表示出来.

解决等级评分问题最常用的方法是将该问题转换为标准的分类问题(*n*-ary classification)[3]或排序(ranking)问题[4].早期的等级评分研究工作曾采用传统文本分类技术来求解,将不同的分数看做是不同的类别.但遇到的最大问题在于,两个相邻分数(如 3 分和 4 分)之间难以区分.在这种情况下,分类技术不能获得令人满意的性能[3],因为传统分类技术擅长于区别差异性较大的类别集,难以有效判别混淆类别[5].为此,一些研究人员[4]采用排序学习(learning to rank)算法来解决等级评分的问题,取得了较好的效果.因此,本文也将研究基于排序技术的等级评分模型,其中采用了一种被广泛使用的排序学习算法,即基于感知机的排序学习算法(perceptron-based ranking,简称 PRanking)[6].该算法也曾被成功应用于餐馆领域的等级评分中[4].

在实际应用中,针对同一评价对象(如餐厅)的不同评论通常是来自于不同的用户,但不同用户对打分标准的理解往往存在不一致性.例如,很难保证不同用户对“很好(5 分)”和“较好(4 分)”两类评价有严格一致性的理解.该不一致问题产生的主要原因是由于在多级评分体系中缺少严格的基准定义,如在 5 分制中如何严格准确定义 4 分与 5 分打分的区别.换句话说,不同用户对不同评分分数所表达程度的理解存在模糊性.例如在餐厅评论方面,每个用户都会从不同的角度(如餐馆的食物或者服务属性)表达他们各自的观点和意见.一般来说,在 5 分评价体系中,用户给出的等级评分 4 分和 5 分通常表达褒义评价,但不同用户给出的 4 分和 5 分在满意度程度上往往存在模糊性.例如,有些较为苛刻的用户给出的 4 分本质上相当于其他较友好用户给出的 5 分所表达的同等极性程度.由此可以得出,几乎不可能保证在不同用户给出的餐馆评论中,从不同的角度能够给出具有一致性评价.通常在等级评分标注具有一致性的训练数据集上,标准机器学习算法性能会比较好.但如果训练集中用户给定的等级评分标注存在不一致性,学习算法性能将会有所下降.为此,本文提出了两种非常简单的方法来改善等级评分模型训练学习过程,以解决用户评价标注存在不一致性的问题.这两种技术均采用优化训练样本选择的方法来改善模型训练学习过程,包括基于容忍度的样本选择方法和基于评价损失的选择方法.

在本文研究的应用任务中,用户经常对评价对象的多个维度(或属性,aspects or features)进行评论,如餐馆的服务维度和价格维度等,而不仅仅简单表达该餐馆总体感觉好或者不好.本文称其为多维度等级评分(multi-aspect rating inference)任务.传统多维度等级评分的解决方案只是将不同维度的等级评分看做是单独的任务,也就是说,不同维度的等级评分模型训练和预测过程相对独立.该传统解决方案比较适合于某些应用任务,其中,不同维度之间不存在明显的相关性.实际上,在餐馆领域的用户评价中,不同维度/属性之间的用户观点评价明显存在一定的关联关系(相关性).例如,从真实餐馆评论中我们发现,当用户对一个豪华餐馆进行评价时,大多数评论中都提到〈环境维度:豪华〉和〈价格:昂贵〉.换句话说,当用户评价餐馆的环境设施比较豪华时,很有可能隐含了该餐馆的价格较贵的评价.如果用户认为该餐馆的服务很差,也会很自然地对餐馆其他方面提供贬义倾向性评价.从上述例子可以看出,豪华的环境维度与昂贵的价格维度之间存在潜在的关联性,这种关联性信息是有助于优化多维度等级评分模型训练和预测性能的.传统的等级评价技术都是每个属性单独进行的.在等级评价过程中,如果考虑用户对评价对象的不同属性之间的评价相关性,将有助于整体性能的改善.我们将研究在评价对象的不同维度之间存在相关性的情况下,如何利用该相关性来优化多维度等级评分模型训练的学习和预测性能.为了达到这个目的,本文提出了面向属性的协同过滤技术(aspect-oriented collaborative filtering,简称 AOCF)

以实现多维度等级评分.在两个公开的英语和汉语餐馆评论数据集上进行验证,实验结果表明,本文提出的优化技术有效地改善了多维度等级评分模型的性能.

1 相关工作

情感分类是与本文研究工作较为相关的研究任务之一,其目标是使用有监督学习的分类技术来分析用户观点的极性,即褒义或贬义[7−9].近几年来,该课题逐渐受到了国内外很多研究人员的关注,并取得了一些研究成果[1−3,10−15].后来,有些研究人员将传统的情感分类任务扩展到等级评分任务,将用户观点文本的极性倾向性采用多值评分来表示[2−4].目前,等级评分通常被视为一个多类/二元分类问题或者排序问题.

Pang 和 Lee[3]曾指出基于内容的等级评分任务存在两个难点:一是用户评分标注不一致的问题,二是等级评分和观点文本之间存在不匹配的问题.但是,目前来看,相关研究都没有深入讨论如何解决等级评分任务中用户评分标注不一致性问题,只是简单假设训练数据中用户标注信息具有一致性.Pang 和 Lee[3]首先在情感分析任务中提出了等级评分问题,并使用了多类文本分类器对用户评论文本进行自动分类,取得了一定效果.Goldberg 和 Zhu[2]采用了基于图的半监督学习算法来解决等级评分问题.这两项研究都集中在单维度(single-aspect)的等级评分任务上.为了解决多维度等级评分问题,Snyder 和 Barzilay[4]提出了基于 PRanking 的联合排序技术.虽然该技术考虑了不同维度之间的相关性,但只是利用了所有维度打分标注一样或不一样两种类型相关性.实际上,用户经常对部分维度(属性)的打分标注一样,如褒性评价或贬性评价,但不能保证对所有维度打分标注一样.在这种情况下,Snyder 和 Barzilay[4]的方法具有很大的局限性,难以有效挖掘任意两个不同维度(属性)的评分标注的关联性.Wang 等人[16]提出了一个潜在维度评分分析(latent aspect rating analysis,简称 LARA)模型,对用户观点评论中潜在维度的等级评分进行分析.实验结果在宾馆评论领域取得一定效果,但没有深入讨论用户评论标注不一致性和多维度协同过滤问题.Zhu 等人[17]提出了维度分割模型来识别具体维度的观点描述文本,用于民意测试(opinion polling)中.但他们的工作只是针对褒贬两类进行分析,并没有针对等级评分任务.本文深入研究了面向多维度等级评分模型两个问题:用户评分标注不一致性和不同维度的用户评分标注相关性.值得一提的是,本文提出的解决技术非常简单、有效.

2 基于内容的等级评分模型

基于内容的等级评分任务可以看做是一个分类问题[3]或排序问题[4].本文中,我们将采用一种广泛使用的排序学习算法——PRanking[6]对基于内容的等级评分进行建模.PRanking 算法本质上通过在线迭代学习机制寻找一个由参数向量 w 和阈值集合 b 所定义的排序规则,并利用该规则来预测输入未标注评论的等级评分.本节将简要介绍一下基于 PRanking 的等级评分模型.关于 PRanking 算法的详细技术细节可参见文献[6].

在不失一般性的前提下,在基于内容的等级评分任务中,给定一个(实例,等级评分)序列$(x_1,y_1),\ldots,(x_t,y_t),\ldots$每个实例(即用户评论文本)$x_t$是一个特征向量,其等级评分 y_t是评分制 Y 中的一个分数,$Y=\{1,2,...,k\}$.基于 PRanking 的等级评分模型目标是学习排序规则 H,其中,H 是从实例到等级评分的映射函数,即 $H:X\to Y$.排序规则 H 基于参数向量 w 和阈值向量 $b=\{b_1,\ldots,b_k\}$进行定义.其中,阈值向量 b 是 k 个阈值的集合,$b_1\leqslant...\leqslant b_{k-1}\leqslant b_k=\infty$.在预测过程中,满足条件 $b_{r-1}<w\cdot x<b_r$ 的测试实例 x 被赋予一个相应的等级评分 $r(\in Y)$.假定给定由 w 和 b 定义的排序规则 $H(\cdot)$,测试实例 x 的预测等级评分函数(排序规则)可以被定义为

$$H(x)=\min_{r\in Y}\{r:w\cdot x-b_r<0\} \tag{1}$$

基于内容的等级评分模型训练目标是学习正确排序规则 $H(\cdot)$,能够保证预测的等级评分和真实的用户标注评分之间的差距达到最小.该差距定义为排序损失(ranking loss),即真实等级评分和预测等级评分之间阈值的个数.本质上,PRanking 算法的学习目标是尽量减少排序损失值.在 PRanking 算法的每一轮训练学习中,一旦发现存在预测错误,将自动更新当前排序规则 $H(\cdot)$,即修改当前的 w 和 b,然后根据预测的等级评分和真实等级评分是否相同来决定给出一个新的排序规则修改决策.

PRanking 算法可以很好地处理等级评分标注具有一致性的训练数据[6].但在真实的评论数据中,往往经常

存在等级评分标注不一致性的问题,在这种情况下,因为不一致性问题可能会导致 PRanking 算法训练学习过程中做出不正确的修正决策,最终导致模型性能的下降.

3 评分标注不一致性问题

由于不同用户对打分机制的理解有所不同,造成评分标注不一致性问题.为了解决这个问题,本节提出了两种优化技术用于改善基于 PRanking 的等级评分模型训练学习过程,最终改善模型预测性能.这两种技术本质上都是通过优化模型训练过程中训练样本的选择来实现的,包括基于容忍度的样本选择方法(tolerance-based selection)和基于排序损失的样本选择方法(ranking-loss-based selection).第 2 节提到,用户评分标注不一致性可能导致 PRanking 学习算法对参数(b 和 w)做出不正确的修正决策.解决该问题的关键在于如何有效地避免 PRanking 算法训练中做出错误参数修正决策.在基于 PRanking 的训练学习过程中,实际上是不可能绝对提前准确知道哪些参数修正决策是正确或错误的.本质上说,本文提出的两种训练样本选取优化技术的目的在于通过引入容忍度或排序损失估计方法,来降低算法学习过程中做出不正确的参数修正决策的几率.下文将详细解释这两种优化技术的基本思想和计算过程.

3.1 基于容忍度的样本选择

本文首先通过一个简单例子来解释不同用户之间评分标注不一致性问题.下面是两个不同用户的真实评论,在食品维度对同一餐馆进行评论:

- 用户 1:因为真的是很经济实惠,超喜欢烤鸡翅~ 等级:4;
- 用户 2:今天我在这里吃晚饭,感觉非常好. 等级:5.

基于上述两个真实用户评论,我们针对 10 个计算机专业硕士研究生进行了一个小规模调查,其中要求每个学生独立地采用 5 分制对上述两个评论进行等级评分标注.从调查结果中发现,80%的学生对两个评论均给出 5 分,剩余 20%的学生对两个评论均给出 4 分.有趣的是,没有一个被测试学生对这两个评论赋予不同的等级分数.需要注意的是,例子中原来两个不同用户给出的等级评分标注是不一样的.

通常情况下,可以合理假设同一个评论者给出的等级评分标注具有一致性.但不同的评论者由于对等级评分机制具有不同的理解,将导致对同一个评论给出的等级评分有可能不同.根据调查结果,上述两个评论应该给出相同的等级评分(5 分或者 4 分).但实际上,用户 1 和用户 2 给出两个不同的打分,这说明在真实数据中存在等级评分标注不一致性问题.

可以设想一下,将上述两个样本评论用于 PRanking 算法的训练学习.假设在第 t 轮训练学习中,如果当前的排序规则 $H(\cdot)$对两个样例评论给出相同的打分(5 分或者 4 分),PRanking 算法会认为至少在一个样本上给出了错误的预测,因此就会自动修改当前排序规则 $H(\cdot)$.换句话说,等级评分标注的不一致性将会导致对当前的排序规则 $H(\cdot)$进行错误的修正操作.为了解决这个问题,本文提出了一种基于容忍度的技术来确定是否需要在当前学习轮次中对当前排序规则进行修正,也称其为基于容忍度的样本选择方法.

在标准的 PRanking 学习算法中,当预测等级评分与真实等级评分不相等时,即做出更新模型排序规则 $H(\cdot)$的决策,称为“硬”约束方式.该判断条件可以定义为

$$\hat{y}^t \neq y^t \tag{2}$$

正如上述例子所讨论的那样,标准 PRanking 学习算法至少对其中一个训练样本做出错误的排序规则 $H(\cdot)$修正决策,这将影响等级评分模型训练学习效果.从真实用户评论数据上可以看出,假设不同用户都想表达褒性观点时,由于每个人对等级评分机制的理解不同,可能会分别给出 4 分或 5 分.但基本上不会给出 1 分或 2 分,因为 1 分或 2 分表示贬性观点.换句话说,如果将等级评分标注转换成为褒贬两类标注的话,所谓的用户标注一致性问题就没有这么严重了.用户对低等级(贬性观点,1 分或 2 分)和高等级(褒性观点,4 分或 5 分)之间的理解和判别是非常清楚的.但在实际情况中,不同用户即使表达相同的褒贬性观点,也非常容易给出不同的等级评分,如 4 分和 5 分表示褒性观点.通常情况下,4 分和 5 分两者的区别较小.基于上述分析,本文提出的基于容忍度的

样本选择方法采用了“软”约束方式来决定是否需要修正当前排序规则 $H(\cdot)$,本质上是优化训练样本选择机制,即将公式(2)修改为

$$|\hat{y}^t - y^t| > \beta \tag{3}$$

其中,β是容忍度因子,取值区间为[1,k–1].k 是当前等级评分机制中最高的分数,如在 5 分制中为 5 分.在 5 级打分机制中,本文实验默认β=1.

3.2 基于排序损失的样本选择

前文提到,PRanking 算法的学习目标是找到能够将排序损失 *RankLoss*(H)达到最小的排序规则 $H(\cdot)$.为此,在学习过程中,直觉上可以直接采用 *RankLoss*(H)作为选择训练实例用于更新当前排序规则的标准,简称为基于排序损失的样本选择方法.基本思想是,如果当更新后排序规则 H^{t+1} 比当前排序规则 H^t 具有更高的排序损失,则保留当前排序规则 H^t 不作任何修正,即取消当前修改决定,并设置 $H^{t+1}=H^t$.基于排序损失的样本选择技术可以形式化定义为

$$\text{IF } \xi(H^{t+1}) > \xi(H^t)+\varepsilon \text{ THEN set } \quad w^{t+1}=w^t, \forall r: b_r^{t+1}=b_r^t \tag{4}$$

其中,$\xi(H)$表示排序规则 $H(\cdot)$的排序损失值,即 *RankLoss*(H);ε是一个固定的常数(0≤ε≤1).在实验中,可以设置ε为 0 或较小的值.

但在实际应用中,基于排序损失的技术存在两个困难:一是计算复杂度的问题.在每一轮学习中,从大规模训练集中计算 *RankLoss*(H)的代价相当高.排序损失的实时计算将导致 PRanking 算法的训练学习过程计算复杂度过高(与标准的 PRanking 算法相比);其次,由于 *RankLoss*(H)是从训练集中估计出来的,具有较低 *RankLoss*(H)的排序规则 $H(\cdot)$不一定能够保证在测试集中具有较好的鲁棒性,甚至会带来较差的测试性能[18].在训练集中学习到的两个不同的排序规则即使具有相同的 *RankLoss*(H),但两者在测试集中的错误率也可能不相同.

为了解决上述这些问题,关键之处是如何在每一轮 PRanking 训练学习中对 *RankLoss*(H)进行有效的估计.本节提出了两种方法来计算 *RankLoss*(H):第 1 种方法是在每轮学习过程中,在相对较小的训练子集中估算 *RankLoss*(H),而不是在整个训练集中进行估算.这样的话,可以将排序损失计算代价减小到最低限度.换句话说,在每一轮训练学习中,*RankLoss*(H)可以有效地从少量的训练数据上估算出来.PRanking 是一种在线学习算法,在第 t 轮学习中,k 个最近使用的实例可以被用来估计当前的排序规则 H^t 和更新后的排序规则 H^{t+1} 的排序损失.虽然从一个小训练集中估计 *RankLoss*(H)的方法可以有效地降低相应的计算代价,但仍无法有效地解决排序规则的鲁棒性问题.另外可行的方法是采用一个规模较小的独立开发集,在此开发集中,可以花费较小的计算成本来估计 *RankLoss*(H).这种方式在很多机器学习任务中用于模型参数优化,以提高机器学习模型的鲁棒性.

另外,在实际应用中,基于排序损失的选择方法很容易得到局部最优解.为了解决这个问题,一种有效的解决办法是首先运行 l 轮标准的 PRanking 训练学习,之后再启动基于排序损失的选择方法(即激活算法的第 5 步).第 l 轮学习得到的排序规则 H^l 作为基于排序损失的选择方法的学习起点.换句话说,在下面所示算法中的步骤 5 在前 l 轮是处于非活动状态.l=0 表示初始的排序规则 H^0 作为基于排序损失的选择方法的训练学习的初始点.

基于 OTSS 的 PRanking 学习算法.

Input: $(x^1,y^1),\ldots,(x^T,y^T)$,排序规则 $H(\cdot)$ (from Eq.1);

Initialize: Set $w^1=0, b_1^1,\ldots,b_{k-1}^1=0, b_k^1=\infty$.

Loop: FOR t=1,2,...,T

1. 取一个新样例 x^t.
2. 根据公式(1)预测它的等级评分 $\hat{y}^t=H(x^t;w^t,b^t)$.
3. 取得样例 x^t 的真实等级评分 y^t.
4. IF $|\hat{y}^t-y^t|>\beta$ THEN 更新 w^t(否则,设置 $w^{t+1}=w^t, \forall r: b_r^{t+1}=b_r^t$,goto Step 1):
   a) FOR r=1,…,k–1: IF $y^t \le r$ THEN $y_r^t=-1$ ELSE $y_r^t=1$.
   b) FOR r=1,…,k–1: IF $(\hat{y}^t-r)y_r^t \le 0$ THEN $\tau_r^t=y_r^t$ ELSE $\tau_r^t=0$.

c) Update $w^{t+1} \leftarrow w^t + \left(\sum_r \tau_r^t\right) x^t$.

d) FOR r=1,…,k−1: update $b_r^{t+1} = b_r^t - \tau_r^t$.

5. IF $\xi(H^{t+1}) > \xi(H^t) + \varepsilon$ THEN set $w^{t+1} = w^t, \forall r : b_r^{t+1} = b_r^t$ (Eq.4).

Output: $H(x;w^{T+1},b^{T+1})$.

4 协同过滤技术

上述两个样本选择优化技术可以用于改善单维度等级评分模型的性能.本节将深入讨论如何有效地考虑评价对象的不同属性之间的关联信息(相关性)以改善多维度等级评分模型.为了解决这个问题,本文采用协同过滤方法(collaborative filtering,简称 CF)来实现多维度等级评分技术.其中,传统的面向用户的协同过滤技术(user- oriented collaborative filtering,简称 UOCF)已被广泛应用于个性化推荐任务中[19,20],例如 Amazon 和 Netflix.其基本思想是充分利用具有相似兴趣的用户的标注结果来帮助预测当前用户的等级评分标注结果.面临的关键技术在于如何挖掘哪些用户可能具有相同的兴趣.但在本文研究的多维度等级评分任务中,例如考虑针对一个餐馆的评论,大部分用户只是提供一个或两个餐馆评论.在这种情况下,由于对每个用户缺少足够的打分标注数据,难以实现不同用户的兴趣相似性计算.前面已经提及,被评价对象(如餐馆)的不同属性(维度)的用户评分标注之间存在一定的相关性,并且挖掘该相关性可以有助于多维度等级评价.为此,本文提出一种面向属性的协同过滤技术(aspect-oriented collaborative filtering,简称 AOCF).其基本思想就是挖掘评价对象的不同属性之间的评分标注相关性(相似性),而不是不同用户之间的兴趣相关性(相似性).为了实现 AOCF 技术,本文采用了一种基于皮尔逊相关系数(Pearson correlation coefficient,简称 PCC)的属性相关性计算方法[20].基于 PCC 的相关性评价被定义为两个归一化打分向量的内积,如下所示:

$$PCC(a,b) = \frac{\mathrm{cov}(a,b)}{\sigma_a \sigma_b} = \frac{\sum_{1 \leqslant i \leqslant m} (a_i - \bar{a})(b_i - \bar{b})}{\sqrt{\sum_{1 \leqslant i \leqslant m} (a_i - \bar{a})^2} \sqrt{\sum_{1 \leqslant i \leqslant m} (b_i - \bar{b})^2}} \tag{5}$$

其中,a={a_1,...,a_m}和 b={b_1,...,b_m}分别表示训练数据中 m 个用户评论中针对属性 a 和 b 的评分标注结果,$\bar{a}$和$\bar{b}$分别表示训练数据中属性 a 和 b 的平均用户等级评分.

定义 Ψ为评价对象的 L 个预定义属性,如服务和环境维度.给定一个未标注文本评论 x,针对属性 a 的等级评分预测可以定义为其他属性的等级评分的加权平均,计算如下:

$$H_{CF}(x,a) = \left[\bar{a} + \frac{\sum_{b \in \Psi \wedge b \neq a} (b_x - \bar{b}) PCC(a,b)}{\sum_{b \in \Psi \wedge b \neq a} PCC(a,b)} \right] \tag{6}$$

其中,$HCF(\cdot)$表示基于协同过滤技术的等级评分预测函数;$[\cdot]$表示取整操作函数,如[4.3]=4 或者[4.6]=5;b_x 表示基于 PRanking 算法的属性 b 的等级评分预测结果.

5 实验分析

5.1 实验设置

本实验首先评价了各种样本选择优化技术在基于内容的等级评分任务中的有效性,包括标准 PRanking 方法(基准系统)、基于容忍度的样本选择方法(TBS)、基于排序损失的选择方法(采用训练子集估计排序损失,RLS-training)、基于排序损失的选择方法(采用开发集估计排序损失,RLS-DevSet)以及 TBS 和 RLS 的组合(TBS-RLS).其中,这种组合方法是使用 TBS 和 RLS-DevSet 技术进行组合而成的.然后,本文进一步评价了面向属性的协同过滤技术(CF)在多维度等级评分任务中的有效性.最后,测试了样本选择优化技术与协同过滤技术相结合的有效性.

在下面的实验设计中,标准 PRanking 算法训练迭代次数(即 T)设为 4.公式(3)中的容忍度因子β设为 1,公式

(4)中的ε设为 0.RLS-Training 方法中的训练子集的大小 k 设为 20.RLS-DevSet 方法中的参数 l 设为 4.基于排序损失的选择方法(RLS)学习过程自动停止条件定义为当无法获得更小的排序损失时.为了测试不同参数设置对模型性能的敏感性,后文的表 8~表 10 显示了 3 个参数 k,l 和ε的不同取值对模型性能的影响分析.

所有实验设计都采用两个公开的餐馆评论方面的数据集,包括英语数据集(EngSet)和中文数据集(ChiSet).英文数据集 EngSet 中包含 4 488 条评论,该数据集曾被用于多维度等级评分任务中[4].数据集 EngSet 中的每条评论均提供了用户在 5 个不同维度,包括食品、服务、价格、环境和体验方面的等级评分标注.每条评论均被表示成由出现至少 3 次以上的词汇和二元组构成的词汇特征向量[4].中文数据集 ChiSet 曾被用于基于多维度的民意调查任务中[21],其中包含针对 100 家餐馆的共 13 350 条中文评论,涉及到环境、食物和服务这 3 个维度的评价.在去除停用词之后,数据集 ChiSet 中的每条评论均被表示为词汇特征向量.上述两个数据集中,等级评分均采用 5 分等级评分制.

为了能够分析不同方法性能的统计意义上的差异性,针对每种被评价方法,所有的实验结果均是 10 次实验的平均值.在每次实验中,使用 80%的数据进行模型训练,10%的数据作为开发集调参,剩余的 10%作为测试集.实验性能采用两种评价指标[4,22],包括排序损失 Ranking-Loss 和 Zero-One 错误率.排序损失值和 Zero-One 值越低,则表明性能越好.

定义测试数据的参考等级评分标注为 $R=(r_1,...,r_n)$,系统预测的等级评分标注为 $R^*=(r_1^*,...,r_n^*)$,排序损失 $loss(R,R^*)$和 Zero-One 错误率 $ZOE(R,R^*)$的计算公式定义如下:

$$loss(R^*,R)=\|R^*-R\|_1/n, ZOE(R^*,R)=|\{i:r_i^*\neq r_i\}|/n.$$

5.2 基于PRanking的等级评分性能

表 1、表 3 和表 5 分别显示了各种基于 PRanking 的等级评分技术在英文和中文数据上的 Ranking Loss 和 Zero-One 性能.同时,表 2、表 4、表 6 和表 7 分别实现了不同技术性能差异性的显著性验证结果.通过解决用户评分标注不一致性问题,基于容忍度的样本选择方法 TBS 有效改善了模型预测性能,即在 EngSet 上取得了较好的平均 Zero-One 性能和在 ChiSet 上取得了较好的平均 Ranking Loss 和 Zero-One 性能.其中,在 EngSet 上的平均 Ranking Loss 性能与基准系统相比没有统计意义上的差异性.从这个实验结果可以看出,ChiSet 上的等级评分标注不一致性问题可能比 EngSet 上要严重得多.

Table 1 Ranking loss on the EngSet for various PRanking-based methods

表 1 不同基于 PRanking 的方法在英文数据集 EngSet 上的排序损失性能

	食物	服务	价值	气氛	体验	Average
PRanking	0.594	0.669	0.721	0.784	0.578	0.669
+TBS	0.582	0.684	0.714	0.789	0.588	0.671
+RLS-Training	0.604	0.677	0.743	0.809	0.605	0.688
+RLS-DevSet	0.577	0.647*	0.699*	0.770*	0.567*	0.652*
+TBS-RLS	0.576*	0.667	0.701	0.778	0.584	0.661

注:带(*)的数字表示最好性能.

Table 2 Paired *t*-tests between various rating inference methods on the EngSet in terms of average ranking loss

表 2 不同等级评分技术在英文数据集上平均排序损失性能的 Paired *t*-test 分析结果

	PRanking	+TBS	+RLS-Training	+RLS-DecSet	+TBS-RLS
PRanking	N/A	~	>>	<<	<<
+TBS	~	N/A	>>	<<	<<
+RLS-Training	<<	<<	N/A	<<	<<
+RLS-DevSet	>>	>>	>>	N/A	>>
+TBS-RLS	>>	>>	>>	<<	N/A

注:在给定 *p-value*>0.05 的前提下,*A*(行)">>"*B*(列),"<<"和"~"分别表示 *A* 比 *B* 好、差和没有统计意义上的性能差异性.

Table 3 Zero-One error on the EngSet for various PRanking-based methods

表 3 不同基于 PRanking 方法在英文数据集 EngSet 上的 Zero-One 错误率性能

	食物	服务	价值	气氛	体验	Average
PRanking	0.444	0.507	0.547	0.578	0.457	0.506
+TBS	0.411	0.485	0.506	0.541	0.437	0.476
+RLS-Training	0.419	0.473*	0.523	0.547	0.452	0.483
+RLS-DevSet	0.425	0.486	0.522	0.557	0.441	0.486
+TBS-RLS	0.401*	0.474	0.495*	0.529*	0.434*	0.467*

注:带(*)的数字表示最好性能.

Table 4 Paired *t*-tests between various rating inference methods on the EngSet in terms of average Zero-One

表 4 不同等级评分技术在英文数据集上平均 Zero-One 性能的 Paired *t*-test 分析结果

	PRanking	+TBS	+RLS-Training	+RLS-DecSet	+TBS-RLS
PRanking	N/A	<<	<<	<<	<<
+TBS	>>	N/A	>>	>>	<<
+RLS-Training	>>	<<	N/A	~	<<
+RLS-DevSet	>>	<<	~	N/A	<<
+TBS-RLS	>>	>>	>>	>>	N/A

注:在给定 *p-value*>0.05 的前提下,*A*(行)">>"*B*(列),"<<"和"~"分别表示 *A* 比 *B* 好、差和没有统计意义上的性能差异性.

Table 5 Ranking loss and Zero-One error on the ChiSet for various PRanking-based methods

表 5 不同基于 PRanking 的方法在中文数据集 ChiSet 上的排序损失率和 Zero-One 错误率性能

	Ranking loss				Zero-One error			
	环境	食物	服务	Average	环境	食物	服务	Average
PRanking	0.885	0.851	0.876	0.871	0.652	0.644	0.646	0.647
+TBS	0.832	0.846	0.820	0.833	0.638	0.644	0.618*	0.633
+RLS-Training	0.928	0.909	0.939	0.925	0.671	0.663	0.658	0.664
+RLS-DevSet	0.772	0.824	0.839	0.811	0.609*	0.636	0.632	0.625
+TBS-RLS	0.764*	0.805*	0.800*	0.790*	0.614	0.628*	0.619	0.620*

注:带(*)的数字表示最好性能.

Table 6 Paired *t*-tests between various rating inference methods on the ChiSet in terms of average ranking loss

表 6 不同等级评分技术在中文数据集上平均排序损失性能的 Paired *t*-test 分析结果

	PRanking	+TBS	+RLS-Training	+RLS-DecSet	+TBS-RLS
PRanking	N/A	<<	>>	<<	<<
+TBS	>>	N/A	>>	<<	<<
+RLS-Training	<<	<<	N/A	<<	<<
+RLS-DevSet	>>	>>	>>	N/A	<<
+TBS-RLS	>>	>>	>>	>>	N/A

注:在给定 *p-value*>0.05 的前提下,*A*(行)">>"*B*(列),"<<"和"~"分别表示 *A* 比 *B* 好、差和没有统计意义上的性能差异性.

Table 7 Paired *t*-tests between various rating inference methods on the ChiSet in terms of average Zero-One

表 7 不同等级评分技术在中文数据集上平均 Zero-One 性能的 Paired *t*-test 分析结果

	PRanking	+TBS	+RLS-Training	+RLS-DecSet	+TBS-RLS
PRanking	N/A	<<	>>	<<	<<
+TBS	>>	N/A	>>	<<	<<
+RLS-Training	<<	<<	N/A	<<	<<
+RLS-DevSet	>>	>>	>>	N/A	<<
+TBS-RLS	>>	>>	>>	>>	N/A

注:在给定 *p-value*>0.05 的前提下,*A*(行)">>"*B*(列),"<<"和"~"分别表示 *A* 比 *B* 好、差和没有统计意义上的性能差异性.

从表 1~表 7 的平均 Ranking Loss 和 Zero-One 性能上可以看出,RLS-DevSet 和 TBS-RLS 能够进一步改善模型的预测性能.其中,TBS-RLS 在 ChiSet 上取得最好的平均 Ranking Loss 和 Zero-One 性能,同时在 EngSet 上也取得了最好的平均 Zero-One 性能.RLS-DevSet 在 EngSet 上取得了最好的平均 Ranking Loss 性能.但 RLS-Training 在两个中英文数据上表现不佳,很多时候表现出负面作用.主要原因在于利用小规模的训练数据来计算当前排序规则的 Ranking Loss,可能会带来训练学习过度拟合问题.从实验训练过程也可看出,RLS-Training 训练学习过程特别容易和快速进入局部最优点,造成模型鲁棒性比较差.

上述实验分析针对的是,平均 Ranking Loss 和 Zero-One 的整体性能.下面简单分析一下各种方法针对具体维度的性能影响.从表 1、表 3 和表 5 中可以看出,RLS-DevSet 和 TBS-RLS 可以改善大多数维度的预测性能,但 TBS-RLS 在 EngSet 的体验维度略有下降.RLS-Training 在很多维度上表现得都不甚理想,最终导致平均性能不理想.TBS 在 EngSet 没有表现出满意的 Ranking Loss 指标评价,但在 Zero-One 指标评价中取得了改善的性能.为了分析不同方法虽然改善了整体性能,但仍对某些具体维度没有取得满意的性能提升的原因,我们深入分析了原始数据.例如,通过对实验数据的分析发现,EngSet 大多数评论中缺乏描述体验维度的文本描述.对于这种情况,实际上,体验维度的预测分析本质上是建立在其他维度的描述文本信息上而得到的.这个问题早先 Pang 和 Lee[3]曾提到过,即某一个维度的等级评分与观点文本内容之间存在不匹配的现象.但他们并没有深入研究这个问题.Zhu 等人[17]曾在民意测验(opinion polling)研究中研究过维度分割模型(aspect-based segmentation),将用户观点评价文本分割成不同具体维度的观点描述,然后进行褒贬性分析.未来,我们将尝试利用该方法来改进等级评分模型.

为了深入分析 3 个参数 k,l 和 ε 的不同取值对等级评分模型学习性能的影响程度,本节在英文数据上进一步设计了一些实验.需要指出的是,上述实验中,设置容忍度因子 β=1.因为在 5 分等级评分机制中,实验结果显示,β>1 的性能非常差,这里不再针对容忍度因子 β 进行敏感性分析.

表 8 显示出 k 的不同取值对 RLS-Training 方法的平均排序损失和 Zero-One 错误率的性能影响.实际上,表中显示的性能差别在 95%置信度水平条件下没有统计意义上的区别.表 8 表明,简单增加用于评估 Ranking Loss 的小训练子集规模,不能保证提高 RLS-Training 方法的性能,并且较大的 k 值将导致模型学习阶段较高的计算代价开销.

RLS-DevSet 方法是基于在线学习算法 PRanking.在这种情况下,RLS-DevSet 学习算法可得益于一个好的学习初始点.其中,参数 l 是用于确定该算法的学习初始点.表 9 显示,l 的不同取值将导致 RLS-DevSet 方法在英文数据上取得不同的性能.其中,l=0 表示初始排序规则 H_0 被用于学习的初始点,其性能表现最差;较大的 l 值将提供较好的学习初始点.类似于 Snyder 和 Barzilay[4]的研究结果,表 9 显示,l=4 能够取得较好的总体性能,并且当 l=5 时并没有取得进一步明显的性能提升.

表 10 显示,默认设置 ε(=0.00)并不是 RLS-DevSet 的最佳性能配置.实际上,通过设置合适的 ε 取值还可以进一步改善 RLS-DevSet 的性能.但在 95%的置信水平条件下,相对于默认配置(ε=0.00),加大取值,如 ε>0.02,并不能得到具有统计意义的性能改善.

Table 8 Sensitive analysis of RLS-Training methods with various k values on the EngSet

表 8 k 的不同取值对 RLS-Training 方法在英文数据上的性能影响分析

k	10	15	20	25	30
RL	0.690	0.682	0.688	0.695	0.688
ZOE	0.481	0.482	0.483	0.485	0.479

其中,k=20 是默认设置.RL 和 ZOE 分别表示排序损失 Ranking-Loss 和 Zero-One 错误率.

Table 9 Sensitive analysis of RLS-DevSet methods with various l values on the EngSet

表 9 l 的不同取值对 RLS-DevSet 方法在英文数据上的性能影响分析

l	0	1	2	3	4	5
RL	0.952	0.735	0.670	0.662	0.652	0.652
ZOE	0.532	0.493	0.488	0.490	0.486	0.490

其中,l=4 是默认设置.RL 和 ZOE 分别表示排序损失 Ranking-Loss 和 Zero-One 错误率.

Table 10 Sensitive analysis of RLS-DevSet methods with various ε values on the EngSet

表 10 ε 的不同取值对 RLS-DevSet 方法在英文数据上的性能影响分析

ε	0.00	0.01	0.02	0.03	0.04	0.05
RL	0.652	0.641	0.650	0.643	0.647	0.654
ZOE	0.486	0.472	0.480	0.485	0.486	0.496

其中,ε=0.00 是默认设置.RL 和 ZOE 分别表示排序损失 Ranking-Loss 和 Zero-One 错误率.

5.3 协同过滤技术性能

本节设计了一些实验比较分析面向属性的协同过滤技术在中英文数据集上的性能.由于基于排序损失的选择方法需要在每轮学习中估计 Ranking Loss,从而导致学习计算代价较高.本节的实验进一步比较、评价了基于容忍度的样本选择方法(TBS)对协同过滤技术的性能影响.

表 11~表 17 显示了面向属性的协同过滤技术(CF)、PRanking 算法(基准系统)和 CF+TBS 在中文和英文两个数据集上的排序损失率和 Zero-One 错误率性能,及其不同技术性能差异性的显著性验证结果.CF+TBS 在两个数据集上都取得了最好的平均排序损失率和 Zero-One 错误率性能.这表明,基于容忍度的样本选择方法(TBS)能够进一步改善协同过滤技术.通过与 PRanking 算法的性能比较分析可以发现:在中文数据集上,协同过滤技术能够进一步改善性能;在英文数据集上,协同过滤技术同样取得了较好的平均排序损失率性能,但是其 Zero-One 错误率性能有细微的下降,本质上没有统计意义上的差异性(95%置信度水平下).通过与表 1~表 3 中显示的 TBS 性能相比较后发现,基于容忍度的样本选择方法能够有效帮助协同过滤技术取得更好的平均排序损失率性能并且能够帮助 PRanking 算法取得较好的平均 Zero-One 错误率性能.协同过滤方法虽然能够取得整体性能的提升,但在某个维度上可能造成负面影响.如同第 5.2 节实验结果分析来看,从实验数据可以看出,用户评论文本中缺乏具体某一维度的观点描述是影响等级评分模型预测的一个关键问题.在这种情况下,基于内容的等级评分模型只能假设整个观点描述文本作为每个维度的描述文本来完成 PRanking 模型训练学习过程.如何自动识别具体维度的观点描述文本片断来训练学习具体维度的等级评分模型,是下一步值得研究的一个问题.

Table 11 Ranking loss on the EngSet for collaborative filtering (CF) methods

表 11 协同过滤方法在英文数据集上的排序损失性能

	食物	服务	价值	气氛	体验	Average
PRanking	0.594	0.669	0.721	0.784	0.578	0.669
CF	0.579	0.662	0.674	0.753	0.591	0.652
CF+TBS	0.565	0.654	0.671	0.763	0.579	**0.646**

注:黑体数字表示最好性能.

Table 12 Paired *t*-tests between various collaborative filtering based rating inference methods on the EngSet in terms of average Ranking Loss

表 12 不同基于协同过滤的等级评分技术在英文数据集上平均排序损失性能的 Paired *t*-test 分析结果

	PRanking	CF	CF+TBS
PRanking	N/A	<<	<<
CF	>>	N/A	<<
CF+TBS	>>	>>	N/A

注:在给定 *p-value*>0.05 的前提下,*A*(行)“>>”*B*(列),“<<”和“~”分别表示 *A* 比 *B* 好、差和没有统计意义上的性能差异性.

Table 13 Zero-One error on the EngSet for collaborative filtering (CF) methods

表 13 协同过滤方法在英文数据集上的 Zero-One 错误率性能

	食物	服务	价值	气氛	体验	Average
PRanking	0.444	0.507	0.547	0.578	0.457	0.506
CF	0.460	0.526	0.531	0.559	0.485	0.512
CF+TBS	0.437	0.501	0.510	0.539	0.464	**0.490**

注:黑体数字表示最好性能.

Table 14 Paired *t*-tests between various rating inference based rating inference methods on the EngSet in terms of average Zero-One

表 14 不同基于协同过滤的等级评分技术在英文数据集上平均 Zero-One 性能 Paired *t*-test 分析结果

	PRanking	CF	CF+TBS
PRanking	N/A	~	<<
CF	~	N/A	<<
CF+TBS	>>	>>	N/A

注:在给定 *p-value*>0.05 的前提下,*A*(行)“>>”*B*(列),“<<”和“~”分别表示 *A* 比 *B* 好、差和没有统计意义上的性能差异性.

Table 15 Ranking loss and Zero-One error on the ChiSet for collaborative filtering (CF) methods

表 15 协同过滤方法在中文数据集 ChiSet 上的排序损失率和 Zero-One 错误率性能

	Ranking loss				Zero-One error			
	环境	食物	服务	Average	环境	食物	服务	Average
PRanking	0.885	0.851	0.876	0.871	0.652	0.644	0.646	0.647
CF	0.845	0.875	0.818	0.846	0.641	0.657	0.631	0.643
CF+TBS	0.827	0.871	0.779	0.826	0.637	0.657	0.613	**0.636**

注:黑体数字表示最好性能.

Table 16 Paired *t*-tests between various collaborative filtering based rating inference methods on the ChiSet in terms of average Ranking Loss

表 16 不同基于协同过滤的等级评分技术在中文数据集上平均排序损失性能的 Paired *t*-test 分析结果

	PRanking	CF	CF+TBS
PRanking	N/A	<<	<<
CF	>>	N/A	<<
CF+TBS	>>	>>	N/A

注:在给定 *p-value*>0.05 的前提下,*A*(行)“>>”*B*(列),“<<”和“~”分别表示 *A* 比 *B* 好、差和没有统计意义上的性能差异性.

Table 17 Paired *t*-tests between various collaborative filtering based rating inference methods on the ChiSet in terms of average Zero-One

表 17 不同基于协同过滤的等级评分技术在中文数据集上平均 Zero-One 性能的 Paired *t*-test 分析结果

	PRanking	CF	CF+TBS
Pranking	N/A	~	<<
CF	~	N/A	<<
CF+TBS	>>	>>	N/A

注:在给定 *p-value*>0.05 的前提下,*A*(行)“>>”*B*(列),“<<”和“~”分别表示 *A* 比 *B* 好、差和没有统计意义上的性能差异性.

6 结论与未来工作

本文主要研究基于内容的多维度等级评分模型中的两个问题:一是训练数据中用户评分标准不一致性的问题.本文提出了两种简单方法来改善基于排序学习的等级评分模型性能.这两种方法都是基于优化训练样本选择技术来优化算法学习过程,包括基于容忍度的样本选择方法和基于评价损失的选择技术;二是如何有效挖掘不同属性之间的关联性来改进多维度等级评分模型.本文提出了一种面向属性的协同过滤技术来改善多维度等级评分性能.在英文和中文两个真实餐馆评论数据集上的实验结果表明,本文提出的方法能够有效改进多维度等级评分模型的预测性能.下一步,我们将进一步重点研究基于内容的等级评分任务中等级评分和观点文本之间不匹配的问题.

References:

[1] Pang B, Lee L. Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, 2008,2(1-2):1−135. [doi: 10.1561/1500000011]

[2] Goldberg AB, Zhu XJ. Seeing stars when there aren't many stars: Graph-based semi-supervised learning for sentiment categorization. In: Rada M, Dragomir R, eds. Proc. of the HLT/NAACL Workshop on TextGraphs. Stroudsburg: Association for Computational Linguistics, 2006. 45−52.

[3] Pang B, Lee L. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In: Proc. of the ACL 2005. 2005. 115−124. [doi: 10.3115/1219840.1219855]

[4] Snyder B, Barzilay R. Multiple aspect ranking using the good grief algorithm. In: Sidner C, Schultz T, eds. Proc. of the NAACL/HLT 2007. Stroudsburg: Association for Computational Linguistics, 2007. 300−307.

[5] Zhu J, Wang H, Zhang X. Confusion class discrimination techniques for text classification. Ruan Jian Xue Bao/Journal of Software, 2008,19(3):630−639 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/19/630.htm [doi: 10.3724/SP.J.1001.2008.00630]

[6] Crammer K, Singer Y. Pranking with ranking. In: Dietterich T, ed. Proc. of the NIPS. 2001. 641−647. http://books.nips.cc

[7] Pang, B, Lee L, Vaithyanathan S. Thumbs up? Sentiment classification using machine learning techniques. In: Proc. of the EMNLP 2002. 2002. 79−86. [doi: 10.3115/1118693.1118704]

[8] Thomas M, Pang B, Lee L. Get out the vote: Determining support or opposition from congressional floor-debate transcripts. In: Jurafsky D, Gaussier E, eds. Proc. of the Conf. on Empirical Methods in Natural Language Processing. Sydney: BPA Digital, 2006. 327−335.

[9] Riloff E, Patwardhan S, Wiebe J. Feature subsumption for opinion analysis. In: Jurafsky D, Gaussier E, eds. Proc. of the Conf. on Empirical Methods in Natural Language Processing. Sydney: BPA Digital, 2006. 440−448.

[10] Huang XJ, Zhao J. Sentiment analysis for Chinese text. Communications of CCF, 2008,4(2):39−47 (in Chinese with English abstract).

[11] Yao TF, Cheng XW, Xu FY, Uszkoreit H, Wang R. A survey of opinion mining for texts. Journal of Chinese Information Processing, 2008,22(3):71−80 (in Chinese with English abstract).

[12] Ni MS, Lin HF. Mining product reviews based on association rule and polar analysis. In: Zhu QM, *et al*., eds. Proc. of the NCIRCS 2007. 2007. 635−642 (in Chinese with English abstract).

[13] Zhao J, Xu HB, Huang XJ, Tan SB, Liu K, Zhang Q. Overview of Chinese opinion analysis evaluation. In: Zhao J, Xu HB, eds. Proc. of the COAE 2008. 2008. 1−22 (in Chinese with English abstract).

[14] Zhao YY, Qin B, Liu T. Sentiment analysis. Ruan Jian Xue Bao/Journal of Software, 2010,21(8):1834−1848 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/3832.htm [doi: 10.3724/SP.J.1001.2010.03832]

[15] Haider S, Mehrotra R. Corporate news classification and valence prediction: A supervised approach. In: Balahur A, Boldrini E, eds. Proc. of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (ACL-HLT 2011). Stroudsburg: Association for Computational Linguistics, 2011. 175−181.

[16] Wang HN, Lu Y, Zhai CX. Latent aspect rating analysis on review text data: A rating regression appraoch. In: Rao B, Krishnapuram B, eds. Proc. of the KDD 2010. 2010. 783−792. [doi: 10.1145/1835804.1835903]

[17] Zhu J, Wang H, Zhu M, Tsou BK, Ma M. Aspect-Based opinion polling from customer reviews. IEEE Trans. on Affective Computing (TAC), 2011,2(1):37−49. [doi: 10.1109/T-AFFC.2011.2]

[18] Su K, Lee C. Speech recognition using weighted HMM and subspace projection approach. IEEE Trans. on Speech and Audio Processing, 1994,2(1):69−79. [doi: 10.1109/89.260336]

[19] Goldberg D, Nichols D, Oki B, Terry D. Using collaborative filtering to weave an information tapestry. Communications of the ACM, 1992,35:61−70. [doi: 10.1145/138859.138867]

[20] Resnick P, Iacovou N, Suchak M, Bergstrom P, Riedl J. Grouplens: An open architecture for collaborative filtering of netnews. In: Smith J, Smith F, eds. Proc. of the ACM Conf. on Computer Supported Cooperative Work. New York: ACM Press, 1994. 175−186.

[21] Zhu JB, Wang HZ, Tsou BK, Zhu MH. Multi-Aspect opinion polling from textual reviews. In: Proc. of the CIKM 2009. 2009. 1799−1802. [doi: 10.1145/1645953.1646233]

[22] Basilico J, Hofmann T. Unifying collaborative and content-based filtering. In: Proc. of the ICML. 2004. 65−72. [doi: 10.1145/1015330.1015394]

附中文参考文献:

[5] 朱靖波,王会珍,张希娟.面向文本分类的混淆类判别技术.软件学报,2008,19(3):630−639. http://www.jos.org.cn/1000-9825/19/630.htm [doi: 10.3724/SP.J.1001.2008.00630]

[10] 黄萱菁,赵军.中文文本情感分析.中国计算机学会通讯,2008,4(2):39−47.

[11] 姚天昉,程希文,徐飞玉,汉思•乌思克尔特,王睿.文本意见挖掘综述.中文信息学报,2008,22(3):71−80.

[12] 倪茂树,林鸿飞.基于关联规则和极性分析的商品评论挖掘.见:朱巧明,编.第 3 届全国信息检索与内容安全学术会议论文集.2007.635−642.

[13] 赵军,许洪波,黄萱菁,谭松波,刘康,张奇.中文倾向性分析评测技术报告.见:赵军,许洪波,编.第 1 届中文倾向性分析评测论文集.2008.1−22.

[14] 赵妍妍,秦兵,刘挺.文本情感分析.软件学报,2010,21(8):1834−1848. http://www.jos.org.cn/1000-9825/3832.htm [doi: 10.3724/SP.J.1001.2010.03832]

王会珍(1980－),女,辽宁建平人,博士,讲师,CCF 会员,主要研究领域为文本分析,机器学习,自然语言处理.

E-mail: wanghuizhen@mail.neu.edu.cn

朱靖波(1973－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为机器翻译,文本分析,机器学习.

E-mail: zhujingbo@mail.neu.edu.cn