软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2017,28(3):676−693 [doi: 10.13328/j.cnki.jos.005154]

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

基于语义约束 LDA 的商品特征和情感词提取*

彭 云[1,2,3], 万常选[1,3], 江腾蛟[1,3], 刘德喜[1,3], 刘喜平[1,3], 廖国琼[1,3]

[1](江西财经大学 信息管理学院,江西 南昌 330013)

[2](江西师范大学 计算机信息工程学院,江西 南昌 330022)

[3](数据与知识工程江西省高校重点实验室(江西财经大学),江西 南昌 330013)

通讯作者: 万常选, E-mail: wanchangxuan@263.net

摘 要: 随着网络购物的发展,Web 上产生了大量的商品评论文本数据,其中蕴含着丰富的评价知识.如何从这些海量评论文本中有效地提取商品特征和情感词,进而获取特征级别的情感倾向,是进行商品评论细粒度情感分析的关键.根据中文商品评论文本的特点,从句法分析、词义理解和语境相关等多角度获取词语间的语义关系,然后将其作为约束知识嵌入到主题模型,提出语义关系约束的主题模型 SRC-LDA(semantic relation constrained LDA),用来实现语义指导下 LDA 的细粒度主题词提取.由于 SRC-LDA 改善了标准 LDA 对于主题词的语义理解和识别能力,从而提高了相同主题下主题词分配的关联度和不同主题下主题词分配的区分度,可以更多地发现细粒度特征词、情感词及其之间的语义关联性.实验结果表明,SRC-LDA 对于细粒度特征和情感词的发现和提取具有较好的效果.

关键词: LDA 模型;语义约束;商品特征;情感词

中图法分类号: TP311

中文引用格式: 彭云,万常选,江腾蛟,刘德喜,刘喜平,廖国琼.基于语义约束 LDA 的商品特征和情感词提取.软件学报,2017,28(3):676−693. http://www.jos.org.cn/1000-9825/5154.htm

英文引用格式: Peng Y, Wan CX, Jiang TJ, Liu DX, Liu XP, Liao GQ. Extracting product aspects and user opinions based on semantic constrained LDA model. Ruan Jian Xue Bao/Journal of Software, 2017,28(3):676−693 (in Chinese). http://www.jos.org.cn/1000-9825/5154.htm

Extracting Product Aspects and User Opinions Based on Semantic Constrained LDA Model

PENG Yun[1,2,3], WAN Chang-Xuan[1,3], JIANG Teng-Jiao[1,3], LIU De-Xi[1,3], LIU Xi-Ping[1,3], LIAO Guo-Qiong[1,3]

[1](School of Information and Technology, Jiangxi University of Finance and Economics, Nanchang 330013, China)

[2](School of Computer and Information Engineering, Jiangxi Normal University, Nanchang 330022, China)

[3](Jiangxi Key Laboratory of Data and Knowledge Engineering (Jiangxi University of Finance and Economics), Nanchang 330013, China)

Abstract: With the development of online shopping, the Web has produced a large quantity of product reviews containing abundant evaluation knowledge about products. How to extract aspect and opinion words from the reviews and further obtain the sentiment polarity of the products at aspect level is the key problems to solve in fine-grained sentiment analysis of product reviews. First, considering certain features of Chinese product reviews, this paper designs methods to derive semantic relationships among words through syntactic analysis, word meaning understanding and context relevance, and then embed them as constrained knowledge into the topic model. Second, a semantic relation constrained topic model called SRC-LDA is proposed to guide the LDA to extract fine-grained topical words. Through the improvement of semantic comprehension and recognition ability of topical words in standard LDA, the proposed model can increase

* 基金项目: 国家自然科学基金(61562032, 61662032, 61662027, 61173146, 61363039, 61363010, 61462037, 61562031); 江西省自然科学基金重大项目(20152ACB20003); 江西省高等学校科技落地计划(KJLD12022, KJLD14035)

Foundation item: National Natural Science Foundation of China (61562032, 61662032, 61662027, 61173146, 61363039, 61363010, 61462037, 61562031); 江西省自然科学基金重大项目(20152ACB20003); 江西省高等学校科技落地计划(KJLD12022, KJLD14035)

收稿时间: 2016-07-03; 修改时间: 2016-09-14; 采用时间: 2016-11-01; jos 在线出版时间: 2016-11-29

CNKI 网络优先出版: 2016-11-29 13:34:56, http://www.cnki.net/kcms/detail/11.2560.TP.20161129.1334.001.html

the words correlation under the same topic and the discrimination under the different topics, thus revealing more fine-grained aspect words, opinion words and their semantic associations. The experimental results show that SRC-LDA is an effective approach for fine-grained aspects and opinion words extraction.

Key words: latent Dirichlet allocation model; semantic constraint; product aspect; opinion word

随着互联网的普及和网络购物所带来的便捷性,网络购物呈现出了前所未有的爆发式增长趋势.由此,在购物网站上产生了大量的商品评论文本数据,且日益呈现大数据化趋势.要从海量的非结构化在线评论文本数据中获得有用的信息,通过人工方式进行处理的难度越来越大,希望通过相应的技术对这些评论文档进行自动化处理、分析,提取有用的知识.在这样的应用需求背景下,出现了针对文本的情感分析(sentiment analysis)技术.情感分析也叫观点挖掘(opinion mining),主要研究人们对某一类实体如产品、服务、事件及其属性所表达的观点、情感和评价的相关问题,情感分析的数据对象主要是文本[1].人们在获取商品总体性评级的同时,有时候还希望了解更细致的商品功能及使用的评价情况,需要进行基于商品特征级别的细粒度的情感分析,以满足人们获取商品局部性特征评价信息的需求.

商品特征是指商品属性及构成商品的各个方面(aspect),包括全局特征和局部特征:全局特征一般指整体对象及其属性,如“这款相机非常不错”中的“相机”“总体质量真的好”中的“质量”;局部特征指整体对象的组成部分及其属性,如“价格很高”中的“价格”“屏幕很清晰”中的“屏幕”.情感词是直接或间接对商品特征进行评价的词语,也有全局情感词和局部情感词之分:全局情感词一般用来描述、评价全局特征,如“相机很好”中的“好”、“质量不错”中的“不错”,且全局情感词具有一定的通用性,有时也可用来修饰局部特征,如“价格不错”等;局部情感词一般用来描述、评价局部特征,如“价格很实惠”中的“实惠”“屏幕很清晰”中的“清晰”.

商品评论是用自然语言表达的非结构化的文本数据,数据量非常庞大,需要综合运用自然语言理解及数据挖掘技术,并有效降低文本的数据表示维度,才有可能实现细粒度的特征和情感词挖掘.利用 LDA 主题模型可以进行文本数据的降维,实现大规模文本数据的主题词提取,并通过主题聚类来获取词语间的关联关系.但 LDA 主题模型偏向于提取高频的全局性主题词和词语共现关系,在主题-词语的概率分配过程中没有考虑词语间的语义关系,导致一些低频的、具有隐含语义关系的特征词和情感词提取的准确率和召回率不高,尤其在具有丰富语义关系的中文商品评论中.具体表现如下.

(1) 难以提取低词频的同义特征.在中文商品评论中,经常会出现多个不同词语描述同一特征,如“价格”“价位”和“价钱”.由于 LDA 模型对高频的“价格”“价钱”较敏感,往往会忽略掉低频的“价位”,从而影响此类特征词的提取率;

(2) 难以发现低词频的情感词.在中文商品评论中,有些情感词只用来修饰某一个或某一类的特征,如“价格很公道”“色彩很鲜艳”中的“公道”“鲜艳”.这类情感词具有一定的专属性,词频相对于全局情感词要低很多,其与特征词的共现关系容易被其他高频情感词所湮没,使得 LDA 模型难以发现这类情感词;

(3) 难以满足细粒度词语的主题分配要求.一篇评论文档往往会对多个不同特征进行评价,如“相机不错,价格很实惠,屏幕清晰,电池也很耐用”中的“相机”“价格”“屏幕”和“电池”,要实现细粒度的特征提取,需要尽量将这些不同特征分配到不同主题;此评论句中也出现了多个情感词,如“不错”“实惠”“清晰”和“耐用”,需要将这些情感词分配到对应其关联特征的不同主题.标准 LDA 倾向于将评论文档中高共现的特征词和情感词分配到同一主题,难以在主题分配中实现细粒度特征和情感词之间的有效区分.

为了解决上述问题,实现细粒度的特征和情感词提取,需要有指导地进行主题词挖掘,即:对主题模型进行约束,形成监督效应来提取符合挖掘目标的主题词.在主题模型中引入 must-link 和 cannot-link 语义约束,使满足 must-link 关系的词语尽量分配到同一主题,而满足 cannot-link 关系的词语尽量分配到不同主题.本文试图从语义关系的发现来探索词语间的关联性,利用关联性进一步对主题模型形成约束机制,从而发现特征和情感词之间的隐含关系.引入词语之间的语义关系可以提升主题模型的语义理解能力,提高识别局部词语间关联关系的能力,更多地发现细粒度的特征和情感词.

本文的主要贡献包括:

(1) 从中文商品评论的语言结构和特点出发,设计了获取特征词-特征词、特征词-情感词和情感词-情感词的 must-link 和 cannot-link 语义关系的方法;

(2) 构建了基于 must-link 和 cannot-link 的语义关系图,设计了利用语义关系图来指导主题模型进行主题-词语分配的约束机制;

(3) 将语义关系知识嵌入到 LDA 模型,提出了细粒度商品特征和情感词提取模型 SRC-LDA.

本文第 1 节介绍相关工作.第 2 节构建语义关系图.第 3 节设计 SRC-LDA 模型.第 4 节进行实验分析.最后部分是总结与展望.

1 相关工作

在商品特征和情感词的提取研究中,主要方法有以下几类.

(1) 基于词频和共现的方法.

在商品特征及情感词的提取中,由于商品特征通常是名词或名词短语,且特征和情感词具有一定共现性,有些研究基于频繁名词和共现规则的方法提取特征和情感词.Hu 等人[2]抽取出现频率大的名词及名词短语作为候选商品特征,通过压缩剪枝和冗余剪枝策略对提取的频繁商品特征进行筛选,抽取特征词附近的形容词作为情感词,再使用关联规则挖掘识别频繁商品特征,最后,利用抽取的情感词来识别非频繁的特征.Popescu 等人[3]将商品特征看作是商品的一部分,使用候选商品特征和领域特征之间的共现来提取商品特征,并使用点互信息 PMI(pointwise mutual information)表示关联程度,最终按关联程度大小选择商品特征.该方法提高了商品特征提取的准确率,但召回率有所下降.

基于词频的方法会造成部分低频特征词的丢失,并容易产生高频的非特征词.同时,随着商品评论数量的增加,共现及关联规则很难覆盖日益复杂的文本语法及语义结构关系.

(2) 基于机器学习的方法.

Jakob 等人[4]利用条件随机场 CRF(conditional random fields)模型提取特征;Jin 等人[5]将特征词和情感词的提取看做是一个序列标注任务:评论中的每个词都对应一个标签类别,提出使用词汇化的隐马尔可夫模型(lexicalized HMM)寻找最有可能的标签序列.Su 等人[6]提出一个相互增强准则来挖掘特征和情感词之间的隐式关联,并基于聚类的方法将隐含特征识别出来.王荣洋等人[7]基于 CRFs 模型研究了多种特征及其组合在特征提取上的效果,重点引入了语义角色标注新特征.

上述基于机器学习的方法需要人工标注数据集,当商品评论的数据量很大时,要耗费大量的人力.

(3) 基于句法依存的方法.

刘鸿宇等人[8]基于句法分析获得名词和名词短语的候选特征,然后,结合 PMI 和名词剪枝算法对候选特征进行筛选获得最终结果.Wu 等人[9]利用依存分析发现评论中商品特征与情感词之间的联系,并使用 Tree-kernel SVM(support vector machine)方法将情感词和商品特征的组合进行分类,分为“相关”“不相关”两类.赵妍妍等人[10]利用统计方法来获取描述评价对象及其评价词语之间修饰关系的句法路径,提出了一种基于句法路径的情感评价单元自动识别方法,并通过句法路径编辑距离的计算来改进情感评价单元抽取的性能.Qiu 等人[11]提出了一种 Double Propagation 方法同时进行情感词和特征词的识别与抽取,在定义一系列种子情感词的基础上,制定了特征词和情感词之间的规则关系,通过不断迭代将情感词抽取与识别出来.姚天昉等人[12]基于依存句法分析总结出“上行路径”和“下行路径”的匹配规则,进而总结出 SBV(主谓关系)极性传递的一些规则,用于情感评价单元的识别.Poria 等人[13]利用商品评论中的语言常识及句法依存树来发现显性和隐性的特征,算法的准确性依赖于句法分析和情感词典.

由于商品评论文本的语法结构较为随意,基于句法依存的方法难以穷尽其句式结构关系,在非规范格式评论文本中难以识别特征和情感词关系.

(4) 基于改进的 LDA 方法.

由于商品评论数据量极大,同时行文较为自由,有些研究者试图利用 LDA(latent dirichlet allocation)主题模

型[14]的文本降维及主题聚类作用,通过提取主题词来发现特征和情感词.LDA 是一种概率生成模型,结构包括 3 层:文档、主题和词语,主要思想是:① 文档是主题的随机混合;② 主题是满足一定概率分布的词语组合.LDA 将表达文本的词向量转化为主题向量,降低了文本维度,同时,在文本生成过程中可以提取主题词.由于 LDA 倾向于产生全局性的主题词,为了提取更多的局部主题词,以下研究对标准 LDA 主题模型进行了扩展,包括两类模型:一类仅提取特征;一类同时提取特征和情感词.

① 特征提取.

Titov 等人[15]将标准 LDA 模型扩展为多粒度 MG-LDA(multi-grain LDA)模型,并假设全局主题倾向于捕获商品总体属性而局部主题倾向于捕获用户评价的商品特征,在此基础上对全局主题和局部主题两类不同类型的主题建模.Andrzejewski 等人[16]将领域知识用 Dirichlet 森林先验的方式加入到 LDA 中,提出了 DF-LDA(dirichlet forest LDA)模型,引入了 Must-Link 和 Cannot-Link 两种约束作为先验知识.但是随着文档数量的增加,该模型的计算复杂度呈指数级增长,给模型的运算带来了困难.Zhai 等人[17]提出了带约束的 LDA (constrained-LDA)模型来实现商品特征抽取及分组,设置了 must-link 和 cannot-link 两种约束类型:一种约束将具有相同成分的特征词归属于同一主题,另一种约束将同一语句中的特征词划分到不同主题(即,一个语句中不会同时出现相同特征的评价).Chen 等人[18]将 must-set 和 cannot-set 引入 LDA,其中,must-set 中的词语属于同一主题的概率较高,而 cannot-set 中的词语属于同一主题的概率较低,提出了 MC-LDA(LDA with m-set and c-set)模型,用于提取特征词.Bagheri 等人[19]提出了基于 LDA 的特征发现模型 ADM-LDA(aspect detection model based on LDA),关注的核心任务是如何从评价句子中提取所需的特征.区别于标准 LDA 的词袋模型,ADM-LDA 假设一个句子中的特征相关词构成一个马尔可夫链,并将这种词语结构信息融入模型;同时,对文档内部特征分布的条件独立性假设进行了松弛处理.马柏樟等人[20]利用 LDA 筛选出候选产品特征词集合,进而通过同义词词林拓展和过滤规则得到最终的产品特征集.Chen 等人[21]在模型中加入先验知识来指导特征提取,提出了 AKL(automated knowledge LDA)模型.先验知识的获取无须人工输入,而是自动从商品评论大数据中得到,并且来自于不同的商品领域.

② 特征和情感词提取.

Lin 等人[22]在标准 LDA 模型的基础上加入了情感层,并考虑每一个情感不同的特征分布,提出了 JST(joint sentiment topic)模型用来同时识别主题和情感.Lu 等人[23]提出了 STM(sentiment topic model)模型,对文档和句子级别的主题联合建模,利用极少量先验知识(种子词形式)来加强主题和特征词的直接关联性,并通过训练总体极性的回归模型进行情感极性预测.Jo 等人[24]假设一个句子仅有一个特征,且句子中的所有词语都由某一个特征来生成,首先提出了 SLDA(sentence-LDA)模型,其主要任务是用来发现特征词;在此基础上提出了 ASUM(aspect and sentiment unification model)模型,它是 SLDA 模型的扩展,将特征和情感合并同时进行建模,用来发现特征词-情感词匹配单元.由于没有特征词和情感词先验关联知识的引入,仅依赖 LDA 本身的先验分布难以识别一些句子级别的词语关系.Moghaddam 等人[25]将评价文本分解为情感短语的形式,提出了 ILDA(interdependent LDA)模型,试图从情感短语中提取特征词及对应的情感词.孙艳等人[26]提出一种无监督的主题情感混合模型 UTSU(unsupervised topic and sentiment unification),通过在标准 LDA 模型中融入情感来实现,可实现文档级别的情感分类.Chen 等人[27]提出了 AMC(automatically generated must-links and cannot-links)主题模型,并在模型中加入了 Must-links 和 Cannot-links 约束知识,用来提取特征词和情感词.Must-links 和 Cannot-links 都是基于已有的 LDA 主题模型从多领域数据中获取,其中,Must-links 中的词语关系知识是利用相同主题下高频率的 top 词语获得,而 Cannot-links 中的词语关系则利用不同主题间的高频率 top 词语获得.由于约束知识的获取直接来源于 LDA,所以会忽略一些低频的特征词和情感词.Dermouche 等人[28]针对目前提取主题词和情感词时往往没有考虑它们之间关联关系的问题,提出了主题-情感 TS(topic-sentiment)主题模型,并基于 Gibbs 抽样过程进行模型参数推导.TS 模型区别于已有模型的特点包括:同样主题的不同描述对应了不同的情感极性,强调情感极性的分布和特定主题的关联性;模型考虑了主题与情感词的关联性,通过在 LDA 中加入情感层来实现,但没有分析主题下特征词和情感词的关联性.欧阳继红等人[29]基于主题情感混合模型 JST 和

R-JST(reverse joint sentiment topic model)并综合文档级和局部级两个粒度上的情感/主题分布,进一步提出了MG-JST(multi grain JST)和 MG-R-JST 模型,能够同时抽取文档的主题和情感信息.一些研究将马尔可夫链、最大熵等引入主题模型,实现特征词、情感词提取以及情感极性分类[30−33].文献[34–36]利用一些外部信息和知识来影响 LDA 的主题词提取,如产品信息、用户行为和人口学知识等.

对 LDA 主题模型的研究现状进行分析,发现 LDA 适于提取全局特征词和全局情感词,难以满足细粒度情感分析的要求,其无监督学习方式也使得提取的主题往往难以符合预期的领域知识挖掘目标.对 LDA 主题模型进行改造,加入先验知识来提高局部主题词的发现率,是目前细粒度情感分析研究的热点和趋势.

LDA 是词袋型概率生成模型,提取的词语关联性主要体现在文档级别的共现,无法深入地理解词语之间的语义关联,从而可能将共现高但无语义关联的词语分配到同一主题,或将共现低但语义关联强的词语分配到不同主题,造成提取的主题词不能真实反映特征和情感词的关系.已有的 must-link 和 cannot-link 语义关系约束获取没有分析特征词和情感词之间的关系,容易造成情感词和特征词的主题分配不准确.如,同一情感词可修饰不同特征、同义情感词可修饰不同特征等.

基于大数据背景下的中文商品评论文本,本文提出了基于特征词和情感词的 3 类 must-link 和 cannot-link 语义关系,在保留 LDA 的大容量文本主题词提取功能的基础上,从语义约束角度对主题模型进行弱监督改造,提升了 LDA 对中文商品评论文本的语义理解能力,使它能够按照预定语义目标进行主题词挖掘,实现细粒度商品特征和情感词的提取.

2 语义关系图构建

引入语义关系的目的是为了影响主题模型的主题-词语分配,通过语义关系尽量发现更多的局部低频特征词和情感词,并增强同类特征及情感词分配到同一主题的概率,同时减少不同类特征及情感词分配到同一主题的概率,提高细粒度主题词及其关系提取的准确率和召回率.词语语义关系的获取来自于文本自身的词、句结构,通过分析候选特征词和候选情感词之间的语义关系,提取特征词之间、特征词和情感词之间以及情感词之间存在的 must-link(w_1,w_2,…,w_n)和 cannot-link(w_1,w_2,…,w_n)关系,其中,w_1,w_2,…,w_n 表示候选的特征词或情感词.属于 must-link 语义关系集合(简记为 MS)的词语应尽量分配到同一主题,而属于 cannot-link 语义关系集合(简记为 CS)的词语应尽量分配到不同主题.

2.1 特征词之间的语义关系获取

(1) 特征词之间的 MS(简记为 MS_{aa})

关注词语的同义性,同义特征词可以互相取代,应尽量分配到同一主题,如“价格”“价钱”和“价位”等.这类词语间具有较强的 must-link 语义关系,一些低频的特征词通过 must-link 关系可关联到高频特征词,从而有利于 LDA 的识别.

候选特征词是名词和动名词,利用《同义词词林扩展版》的层级结构可以获取候选特征词之间的同义关系,见公式(1).

$$S(w_i,w_j)=\begin{cases}1, \text{ if } w_i \text{ and } w_j \text{ have same } l_{1\sim4}\\ 0, \text{ else}\end{cases} \tag{1}$$

其中,$S(w_i,w_j)$等于 1 表示词语 w_1 和 w_2 具有同义性,$l_{1\sim4}$ 表示同义词词林的前 4 层结构.

在获取 MS_{aa} 候选特征词后,可构建如图 1 所示的语义关系图,每一个连通子图对应一个同义特征词聚类簇.

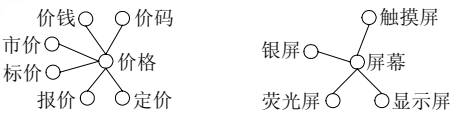


Fig.1 Semantic relationships diagram of must-link between aspects and aspects

图 1 特征词-特征词的 MS_{aa} 语义关系图

(2) 特征词之间的 CS(简记为 CS_{aa})

考虑同一句子中特征词的不可重复性,即,一个复句中多个单句的评价特征的互斥性.

例 1:“板板不错,外观很漂亮,价格合适,图像清晰,性能还行.”的词性标注和依存句法分析如图 2 所示,其中,“板板” “外观” “价格” “图像”和“性能”是 5 个不同的特征.考虑到复句中的候选特征之间具有一定的句法依存关系,设置句法规则来获取候选特征词.

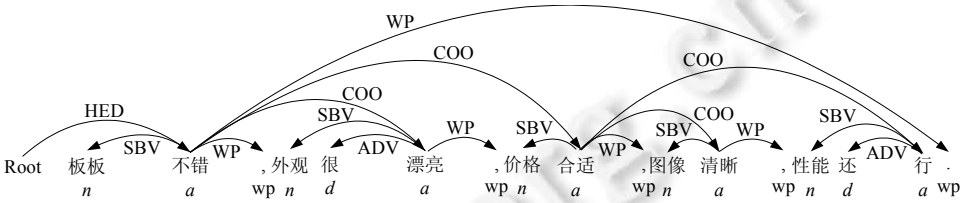


Fig.2 POS tagging and dependency parsing of Exp.1

图 2 例 1 的词性标注和依存句法分析

规则 1. 一个复句中的单句满足 SBV(主谓关系)依存结构关系,对应的主语名词(或动名词)组成候选特征词集.

根据规则 1,从图 2 中可以获得候选特征词集{板板,外观,价格,图像,性能}.

在获取 CS_{aa} 的多个候选特征词集后,可以进行集合间的合并,使得各集合之间不存在共有词语,可构成如图 3 所示的语义关系图,其中,词语节点间有连接边表示存在 cannot-link 关系.

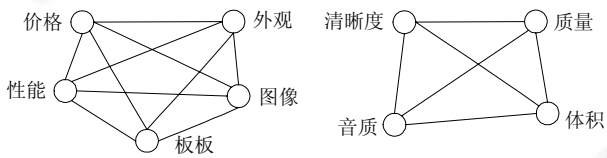


Fig.3 Semantic relationships diagram of cannot-link between aspects and aspects

图 3 特征词-特征词的 CS_{aa} 语义关系图

2.2 特征词和情感词之间的语义关系获取

(1) 特征词和情感词之间的 MS(简记为 MS_{ao})

不考虑 LDA 容易发现的高频全局共现关系,主要关注局部特征词和局部情感词之间的共现关系,尤其是情感词修饰特征词的专有关系.

例 2:“价格很公道”“霸气的外观”的词性标注和依存句法分析如图 4 所示,其中,情感词“公道”“霸气”只修饰一个或一类特征词,这些情感词一般词频较低,不易被 LDA 发现.通过句法结构分析和词性关系规则识别候选特征词和情感词,并利用改进的 PMI 算法进行共现的专有性识别,实现候选特征词和情感词的筛选.

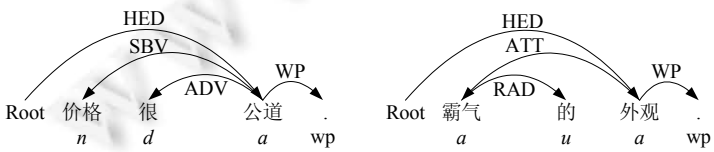


Fig.4 POS tagging and dependency parsing of Exp.2

图 4 例 2 的词性标注和依存句法分析

规则 2. 一个单句中满足 SBV(主谓关系)或 ATT(定中关系)依存结构关系,对应的名词(或动名词)为候选特征词,对应的形容词为候选情感词.

根据规则 2,从图 4 中可以识别候选特征词和情感词集{(价格,公道),(外观,霸气)}.

在句法分析的基础上,设计改进的 PMI 算法来计算特征词和情感词之间的关系,以获取符合语义要求的候选特征词和情感词.计算候选特征词和情感词之间的语义关系强度 OES-PMI(opinion exclusive in sentence PMI)见公式(2),即使共现频率不高,但情感词对于特征词具有专属性,也会有较高的语义关系强度.

$$OSE-PMI(w_i,w_j)=\left|\frac{\lg f_c(w_i,w_j)}{\lg f(w_i)\lg(f(w_j)-f_c(w_i,w_j))}\right|, f(w_i)<\xi_1, f(w_j)<\xi_2 \tag{2}$$

其中,ξ_1 是局部特征词频率阈值,ξ_2 是局部情感词频率阈值,$f(w_i)$是候选特征词 w_i 的词频,$f(w_j)$是候选特征词 w_j 的词频,$f_c(w_i,w_j)$是候选特征词 w_i 和候选情感词 w_j 以句子为单位的共现频率.取关系值大于一定阈值的候选特征词和情感词匹配单元,构成特征词-情感词集.

在获取 MS_{ao} 候选词语后,可构建如图 5 所示的语义关系图,每一个连通子图对应一个候选特征词及多个相关情感词,其中,虚线表示被筛除的关系.

Fig.5 Semantic relationships diagram of must-link between aspects and opinions

图 5 特征词-情感词的 MS_{ao} 语义关系图

(2) 特征词和情感词之间的 CS(简记为 CS_{ao})

考察一个复句中评价多个特征的时候,局部情感词对于修饰的特征往往具有专有性,这类情感词和其他特征词具有 cannot-link 关系.如复句“像素高,内存大,运行速度快.”,其单句中的情感词和其他单句的特征词具有一定排斥性,如,“高”就不能用来评价“内存”及“运行速度”.

特征词和情感词之间的 CS_{ao} 可以结合 CS_{aa} 和 MS_{ao} 进行获取,如“像素高,内存大,运行速度快.”中的 CS_{aa} 是{像素,内存,运行速度},其对应的 MS_{ao} 是{(像素,高),(内存,大),(运行速度,快)},其候选特征词和情感词语义关系构建如图 6 所示.

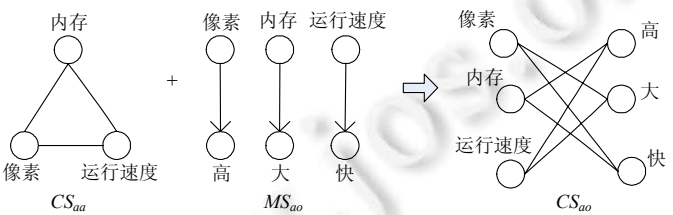


Fig.6 Semantic relationships diagram of cannot-link between aspects and opinions

图 6 特征词-情感词的 CS_{ao} 语义关系图

2.3 情感词和情感词之间的语义关系获取

(1) 情感词和情感词之间的 MS(简记为 MS_{oo})

考虑评价同一特征的近义词及反义词,这些词语之间具有 must-link 关系,如“图像很清晰”“图像很清楚”和“图像很模糊”中的{清晰,清楚,模糊}.通过近义词及反义词,可以发现一些低频的修饰同一特征的情感词.

情感词和情感词之间的 MS_{oo} 获取可以结合 MS_{ao} 和近、反义词计算进行,近、反义词的计算利用同义词词林,公式(3)中,$SA(w_i,w_j)$等于 1 表示词语 w_1 和 w_2 是近义词,等于−1 表示词语 w_1 和 w_2 是反义词;$l_{1\sim4}$ 表示同义词词林的前 4 层结构,l_4 表示同义词词林的第 4 层结构.

$$SA(w_i,w_j)=\begin{cases}1, & \text{if } w_i,w_j\in same\ l_{1\sim4}\\ -1, & \text{if } w_i,w_j\in same\ l_{1\sim3}\wedge w_i,w_j\notin same\ l_4\end{cases} \tag{3}$$

构建 MS_{oo} 候选情感词的语义关系如图 7 所示,通过 MS_{ao} 可以发现候选特征词“外观”对应的情感词“漂亮”,利用近、反义词计算可获得情感词集{美丽,丑陋,难看},这些情感词之间具有 must-link 关系.

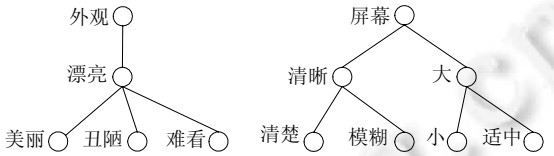


Fig.7 Semantic relationships diagram of must-link between opinions and opinions

图 7 情感词-情感词的 MS_{oo} 语义关系图

(2) 情感词和情感词之间的 CS(简记为 CS_{oo})

考虑复句中不同单句的局部情感词和其他单句局部情感词的关系,如,“相机不错,价格很便宜,外观很时尚,物流很快,服务很周到.”中的局部情感词“便宜”“时尚”“很快”和“周到”之间具有互斥性.

要获取情感词和情感词之间的 CS_{oo},首先利用 MS_{ao} 获得复句中单句的局部情感词,这些情感词之间形成 cannot-link 关系,其语义关系如图 8 所示,其中,“不错”是被筛除的情感词.

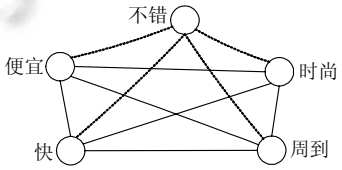


Fig.8 Semantic relationships diagram of cannot-link between opinions and opinions

图 8 情感词-情感词的 CS_{oo} 语义关系图

2.4 语义关系图的融合

对以上语义关系图进行融合,可以构建语义关系图 MSG(must-link semantic graph)和 CSG(cannot-link semantic graph),分别如图 9、图 10 所示.

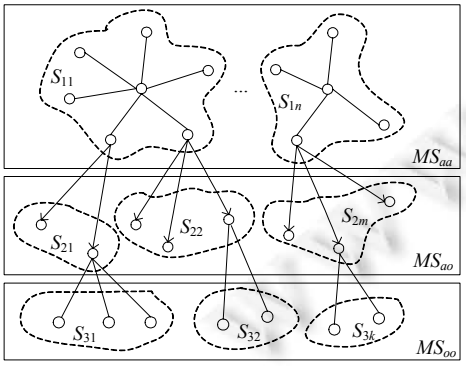


Fig.9 Semantic relationships of *MSG*

图 9 语义关系图 *MSG*

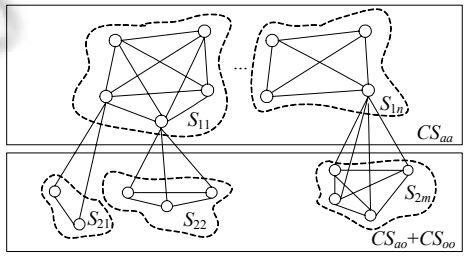


Fig.10 Semantic relationships of *CSG*

图 10 语义关系图 *CSG*

在图 9 中,包括 MS_{aa},MS_{ao} 和 MS_{oo} 这 3 层.其中,MS_{aa} 中包含多个不相交的同义特征词集合;MS_{ao} 层中的局部情感词集合关联于对应的特征词;MS_{oo} 中的情感词和 MS_{ao} 层中情感词形成近义、反义关系,并同时关联到对应

的特征词.在图 10 中,包括 CS_{aa},CS_{ao}+CS_{oo} 这两层.其中,CS_{aa} 中包含多个不相交的特征词集合;在 CS_{ao}+CS_{oo} 层中,首先利用 CS_{ao} 获取局部情感词和对应特征词的 cannot-link 关系,然后利用 CS_{oo} 得到对应于同一特征词的局部情感词之间的 cannot-link 关系.

3 SRC-LDA 模型设计

3.1 语义约束机制

语义约束可以增加相同主题下词语的语义一致性,同时减少不同主题下词语之间的耦合性,从而提取更多的细粒度特征词和情感词.主要从以下 3 个方面考虑约束机制的设计.

(1) 改善 LDA 模型的语义理解能力,减弱无语义相关共现关系的影响,尽可能多地发现符合局部语义关系的特征和情感词;

(2) 在主题模型的主题-词语分配中,增强满足 MSG 关系词语的同一主题分布概率,减弱满足 CSG 关系的词语同一主题的分布概率,提高同主题词语间的内聚度,减少不同主题词语间的耦合度;

(3) 语义约束要弥补 LDA 对于低频关系识别的不足,提高共现频率低、但具有明显特征和情感词语义关系的词语的分配权重,更多地发现低频隐含关系.

将语义约束知识加入到 LDA,通过概率增益对主题-词语的分配产生影响,分两种情况进行计算.

(1) 在对名词或动名词 w 进行主题分配时,考察其是否属于 MSG 和 CSG:如不属于,则不进行分配约束;否则,按照公式(4)进行分配概率增益 $q^k(w)$的计算,即,计算语义约束对于 w 分配到主题 k 的影响值.

$$q^k(w)=\lambda_a n_{mk}-(1-\lambda_a)n_{ck} \tag{4}$$

其中,λ_a 是分配调节因子,n_{mk} 是主题 k 中已分配词语在 MSG 中的个数,n_{ck} 是主题 k 中已分配词语在 CSG 中的个数.

(2) 在对形容词 w 进行主题分配时,分两种情况计算概率增益 $q^k(w)$:以单句为单位考察和其相邻的名词或动名词 w'是否同属于 MSG.

- 若属于则增强 w 和 w'属于同一个主题的概率,$q^k(w)$计算如公式(5),其中,η^o 是分配系数.

$$q^k(w)=\lambda_o \eta^o-(1-\lambda_o)n_{ck} \tag{5}$$

- 否则,$q^k(w)$计算见公式(6).

$$q^k(w)=\lambda'_o n_{mk}-(1-\lambda'_o)n_{ck} \tag{6}$$

对公式(4)~公式(6)进行归一化,得到值小于 1 的 $q^k(w)$.

约束机制的加入,在一定程度上指导主题-词语的概率分配,同时又保留了 LDA 本身的主题-词语分配机制,在发挥其主题聚类作用的同时,实现细粒度的主题词挖掘,即细粒度特征和情感词的提取.

3.2 SRC-LDA模型结构

将上述约束加入到 LDA 形成 SRC-LDA(semantic relation constrained LDA)模型,如图 11 所示,图中的符号说明见表 1.

Table 1 Notations in SRC-LDA model

表 1 SRC-LDA 模型符号说明

符号	说明	符号	说明
α	文档-主题分布的 Dirichlet 参数	φ_m	must-link 约束的主题-词语分布
β	主题-词语分布的 Dirichlet 参数	S_{CSG}	cannot-link 语义关系集
θ	文档-主题分布	φ_c	Cannot-link 约束的主题-词语分布
φ	主题-词语分布	T	主题个数
w	词语	M	文档个数
w'	w 的相邻词语	N_s	文档的句子个数
z	主题	N	句子的词语个数
S_{MSG}	must-link 语义关系集		

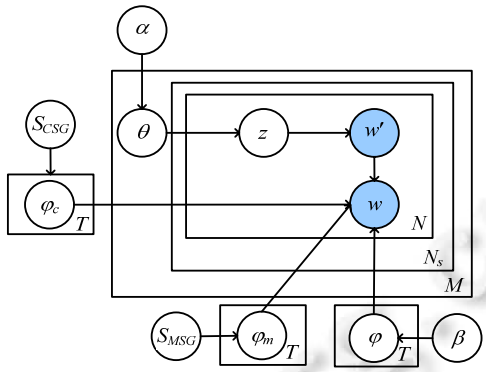


Fig.11 SRC-LDA model

图 11 SRC-LDA 模型

3.3 文档生成过程

SRC-LDA 模型的文档生成过程如下.

- 选择主题分布 θ~*Dirichlet*(α);
- 选择词语分布 φ:
  - if ($w_i \in S_{MS}$) 选择词语分布 $\varphi_m \sim \zeta^m \cdot Dirichlet(\beta)$($\zeta^m$ 是 must-link 因子);
  - else if ($w_i \in S_{CS}$) 选择词语分布 $\varphi_c \sim \zeta^c \cdot Dirichlet(\beta)$($\zeta^c$ 是 cannot-link 因子);
  - else 选择词语分布 $\varphi \sim Dirichlet(\beta)$;
- 对于文档 d 的句子 s_k 中的词语 w_i:
  - 选择主题 $z_j \sim \theta$;
  - if ($w_i \in S_{MSG}$) 选择词语 $w_i \sim \varphi_m$;
  - if ($w_i \in S_{CSG}$) 选择词语 $w_i \sim \varphi_c$;
  - else 选择词语 $w_i \sim \varphi$.

3.4 模型参数计算

SRC-LDA 在原有的 LDA 模型基础上添加了语义约束条件.在计算词语 w 属于某主题 z 的分布概率之前,先进行两类条件分析.

① 词性判断;

② 词语 w 的语义关系判断.

模型参数的计算主要包括文档-主题分布 θ和主题-词语分布 φ的计算,要实现这两个参数的计算,首先要对 Gibbs 抽样的概率公式进行求解.SRC-LDA 模型在标准 LDA 的基础上引入了约束变量 γ,Gibbs 抽样的概率公式见公式(7).

$$P(z_i = k \mid \boldsymbol{w}, \boldsymbol{z}_{-i}, \alpha, \beta, \gamma) = \frac{P(\boldsymbol{w}, \boldsymbol{z}, \alpha, \beta, \gamma)}{P(\boldsymbol{w}, \boldsymbol{z}_{-i}, \alpha, \beta, \gamma)} \propto \frac{P(\boldsymbol{w}, \boldsymbol{z}, \alpha, \beta, \gamma)}{P(\boldsymbol{w}_{-i}, \boldsymbol{z}_{-i}, \alpha, \beta, \gamma)} \tag{7}$$

由于

$$\frac{P(\boldsymbol{w}, \boldsymbol{z}, \alpha, \beta, \gamma)}{P(\boldsymbol{w}_{-i}, \boldsymbol{z}_{-i}, \alpha, \beta, \gamma)} = \frac{P(\boldsymbol{w}, \boldsymbol{z} \mid \alpha, \beta, \gamma)}{P(\boldsymbol{w}_{-i}, \boldsymbol{z}_{-i} \mid \alpha, \beta, \gamma)} \tag{8}$$

由公式(7)、公式(8)可以得到公式(9).

$$P(z_i = k \mid \boldsymbol{w}, \boldsymbol{z}_{-i}, \alpha, \beta, \gamma) \propto \frac{P(\boldsymbol{w}, \boldsymbol{z} \mid \alpha, \beta, \gamma)}{P(\boldsymbol{w}_{-i}, \boldsymbol{z}_{-i} \mid \alpha, \beta, \gamma)} = P(w_i, z_i = k \mid \alpha, \beta, \gamma) \tag{9}$$

公式(9)可进一步展开为公式(10).

$$\begin{aligned}P(w_i,z_i=k\mid\alpha,\beta,\gamma)&=P(w_i\mid z_i=k,\alpha,\beta,\gamma)P(z_i=k\mid\alpha,\beta,\gamma)\\&=P(w_i\mid z_i=k,\beta,\gamma)P(z_i=k\mid\alpha)\\&=P(w_i\mid z_i=k,\beta)P(w_i\mid z_i=k,\gamma)P(z_i=k\mid\alpha)\end{aligned}\tag{10}$$

由公式(9)、公式(10)可得公式(11).

$$P(z_i=k|\boldsymbol{w},\boldsymbol{z}_{-i},\alpha,\beta,\gamma)\propto P(w_i|z_i=k,\beta)P(w_i|z_i=k,\gamma)P(z_i=k|\alpha)\tag{11}$$

最终,SRC-LDA 模型 Gibbs 抽样的概率估算公式为

$$P(z_i\mid\boldsymbol{w},\boldsymbol{z}_{-i},\alpha,\beta,\gamma)\propto P(w_i\mid z_i,\gamma)\frac{\{N_{w_i,k}\}_{-i}+\beta}{\{N_k\}_{-i}+V\beta}\frac{\{N_{d,k}\}_{-i}+\alpha}{\{N_d\}_{-i}+T\alpha}\tag{12}$$

其中,$\{N_{w_i,k}\}_{-i}$表示词语 w_i 属于主题 k 的次数(本次抽样除外),$\{N_k\}_{-i}$表示所有词语属于主题 k 的次数(本次抽样除外),V 是词语总数,$\{N_{d,k}\}_{-i}$表示 w_i 所在文档 d 属于主题 k 的次数(本次抽样除外),$\{N_d\}_{-i}$表示 w_i 所在文档 d 属于所有主题的次数(本次抽样除外).

由公式(12)得到分布参数θ和φ的计算公式,见公式(13)、公式(14),其分布对应于文档 d、主题 k 和词语 w.

$$\theta_{k,d}=\frac{\{N_{d,k}\}_{-i}+\alpha}{\{N_d\}_{-i}+T\alpha}=\frac{C_{dk}^{DK}+\alpha}{\sum_{k_{-1}}C_{dk}^{DK}+T\alpha}\tag{13}$$

$$\varphi_{w,k}=P(w_i\mid z_i=k,\gamma)\frac{\{N_{w_i,k}\}_{-i}+\beta}{\{N_k\}_{-i}+V\beta}=(1+q^k(w))\frac{C_{wk}^{WK}+\beta}{\sum_{w_{-1}}C_{wk}^{WK}+V\beta}\tag{14}$$

其中,C_{dk}^{DK} 是文档 d 在主题 k 中的出现次数,C_{wk}^{WK} 是词语 w 在主题 k 中的出现次数,k_{-1}表示本次抽样主题 k 外的所有主题,w_{-1}表示词语集合中除本次抽样词语 w 外的所有词语,$q^k(w)$是词语 w 的语义约束产生的概率增益.

4 实验分析

4.1 数据集选择及设置

数据集采集于淘宝(www.taobao.com)、天猫(www.tmall.com)和京东商城(www.jd.com)用户的评论数据,利用爬虫软件设置关键字“平板电脑”共采集了 222 507 篇评论文档.为了避免评论文档字数太少而影响可信度,剔除了少于 10 个字的评论文档,得到 189 416 篇评论文档,共包含 1 048 530 个句子.从原始评论文档中提取的 *MSG* 包含 63 个 must-link 集合,CSG 包含 167 个 cannot-link 集.分词工具采用中国科学院的 ICTCLAS,依存句法分析采用哈尔滨工业大学的 LTP[37].

为了验证本文提出的 SRC-LDA 主题模型的效果,分别选取 LDA[14],ASUM[24],AMC[27]和 TS[28]等主题模型进行实验效果对比分析,其中,ASUM,AMC 和 TS 都是较典型的情感分析模型,并且 SRC-LDA 模型与 AMC 模型更为接近.实验文本保留词性为名词、动名词和形容词的词语,均采用 Gibbs 抽样进行参数估计.主题模型测试集和训练集评价文档数的比例设置为 1:10.相关系数设置为:文档-主题概率分布参数α为 50/K,K 为主题个数;top-n 取值为 10(即:在每个主题中,取按概率降序排列的前 top-n 个词语作为主题词);主题-词语概率分布参数β为 0.01;抽样次数为 1 000 次,采用 10-fold 交叉验证.

4.2 评价标准

采用人工方式标注平板电脑评论数据中的特征词集、情感词集及特征词-情感词评价单元集.标注结果统计如下.

(1) 特征词集有 266 个特征词,其中,全局特征词 27 个,局部特征词 239 个;

(2) 情感词集有 156 个情感词,其中,全局情感词 45 个,局部情感词 111 个;

(3) 特征词-情感词评价单元集有 301 个评价单元,其中,全局特征词-情感词评价单元有 58 个,局部特征词-情感词评价单元有 243 个.

以人工标注的数据作为基准对实验结果进行评价,采用准确率(precision)、召回率(recall)来评估 3 个模型的特征词、情感词及局部特征词-情感词评价单元提取效果.

准确率计算为

$$Pa=\frac{Na}{\sum_{i=1}^{K}N_{n,vn}^{i}@\text{top-}n},Po=\frac{No}{\sum_{i=1}^{K}N_{a}^{i}@\text{top-}n},P_p=\frac{N_p}{K\cdot\text{top-}n} \tag{15}$$

其中,P_a,P_o,P_p 分别是特征词、情感词和词语关联组合的准确率,N_a,N_o,N_p 分别是准确提取的特征词、情感词和局部特征词-情感词评价单元的个数,$N_{n,vn}^{i}@\text{top-}n$ 是主题 i 下 top-n 个词语中名词及动名词个数,$N_a^i@\text{top-}n$ 是主题 i 下 top-n 个词语中形容词个数,K 是主题个数.

召回率计算为

$$R=\frac{\sum_{i=1}^{K}N_i@\text{top-}n}{N_s} \tag{16}$$

其中,$N_i@\text{top-}n$ 是主题 i 下 top-n 个词语中提取的不重复的准确特征词、情感词或局部特征词-情感词评价单元个数,N_s 是人工标注的对应词语个数.

4.3 不同模型的比较分析

1) 特征词提取

特征词提取的准确率和召回率如图 12 所示,其中,横坐标为主题个数 K,纵坐标为准确率 P 或召回率 R.

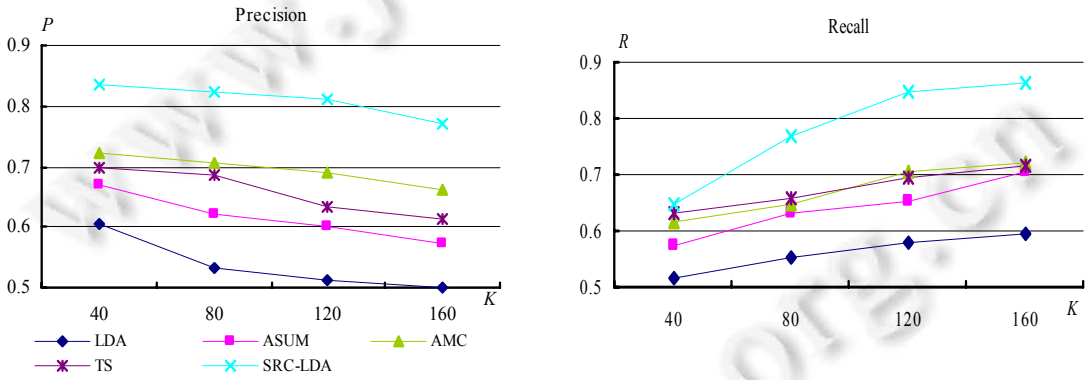


Fig.12 Precision and recall of aspect extraction

图 12 特征词提取的准确率和召回率

从图 12 可以看出:SRC-LDA 模型相对于其他模型,在不同主题数 K 时都具有更高的准确率 P 和召回率 R;而 LDA 模型的 P 和 R 比其他模型都低;与 SRC-LDA 模型差异最小的是 AMC 模型.

关于准确率 P 的差异情况分析如下.

(1) 当 K=160 时,SRC-LDA 和 AMC 的 P 值相差最小,相差 11.1 个百分点;

(2) 当 K=120 时,SRC-LDA 和 AMC 的 P 值相差最大,相差 12.5 个百分点;

(3) SRC-LDA 比 AMC 的 P 值平均高出 11.7 个百分点.

原因分析:由于 LDA 模型偏向于提取全局特征词,随着主题数目的增加,难以识别更多的低频局部特征词,其准确率值下降较快且明显低于其他模型;ASUM 和 TS 考虑了主题词和情感词的关系,提高了特征识别的准确率;由于 AMC 中引入了基于 LDA 的 must-links 和 cannot-links 约束,改善了主题对特征词的识别效果,准确率仅低于 SRC-LDA;SRC-LDA 加入了基于语义的 must-link 和 cannot-link 约束,在词语的主题分配中增加了低频同义特征词的分配概率增益,同时减少了不同类特征词之间的主题分配干扰,尤其在局部特征词的识别上具有明显优势.

关于召回率 R 的差异情况分析如下.

(1) 当 K=40 时,SRC-LDA 和 AMC 的 R 值相差最小,相差 3.2 个百分点;

(2) 当 K=120 时,SRC-LDA 和 AMC 的 R 值相差最大,相差 14.2 个百分点;

(3) SRC-LDA 比 AMC 的 R 值平均高出 11 个百分点.

原因分析:从召回率值的变化来看,当主题数较低时,3 个模型差异不大的原因在于主题词总数偏低,提取的一般是全局特征词;在主题数逐渐增加的时候,SRC-LDA 的召回率具有较明显的优势,表明对低频特征词有较好的识别度,而其他模型难以进一步发现低频特征词.例如,SRC-LDA 可以发现低频的同义特征词,如与特征词“价格”同义的低频特征词“售价”“卖价”和“现价”等,从而提高了召回率.

2) 情感词提取

情感词提取的准确率、召回率如图 13 所示,其中,横坐标为主题个数 K,纵坐标为准确率 P 或召回率 K.

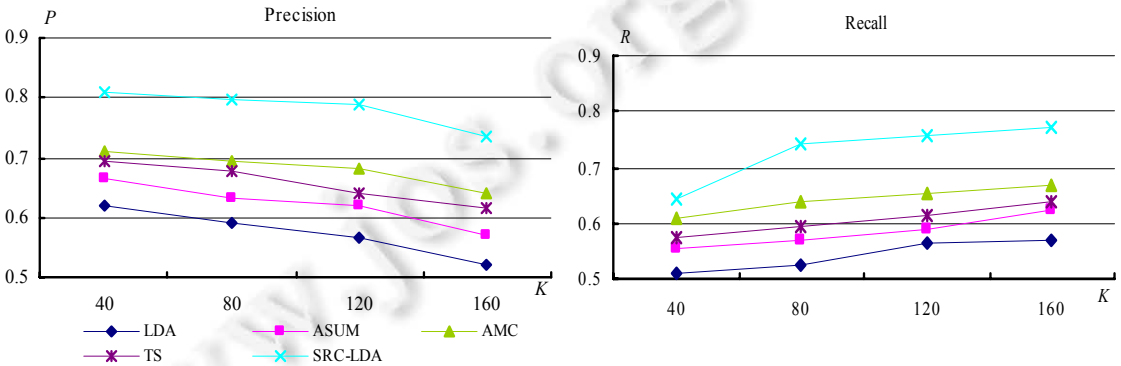


Fig.13 Precision and recall of opinion extraction

图 13 情感词提取的准确率和召回率

从图 13 可以看出:SRC-LDA 的准确率优势较明显,召回率优势随着 K 的增加也逐渐表现出来.

关于准确率 P 的差异情况分析如下.

(1) 当 K=160 时,SRC-LDA 和 AMC 的 P 值相差最小,相差 9.5 个百分点;

(2) 当 K=120 时,SRC-LDA 和 AMC 的 P 值相差最大,相差 10.6 个百分点;

(3) SRC-LDA 比 AMC 的 P 值平均高出 10.1 个百分点.

原因分析:由于 LDA 模型偏向于提取和全局特征词关联较多的全局情感词,难以识别更多的中、低频情感词,其准确率值明显低于其他模型;ASUM 和 TS 模型关注了情感词和主题词的关联性,其准确率值均高于 LDA;AMC 中引入了基于 LDA 的 must-links 和 cannot-links 约束,在一定程度上改善了主题对情感词的聚集度和区分度,准确率仅低于 SRC-LDA;SRC-LDA 中的 must-link 和 cannot-link 约束,在情感词的主题分配中增加了低频情感词关联到特征词的概率增益,同时减少了不同类情感词之间的主题分配干扰,在局部情感词的识别上具有明显优势.

关于召回率 R 的差异情况分析如下.

(1) 当 K=40 时,SRC-LDA 和 AMC 的 R 值相差最小,相差 3.5 个百分点;

(2) 当 K=120 和 K=160 时,SRC-LDA 和 AMC 的 R 值相差最大,相差 10.7 个百分点;

(3) SRC-LDA 比 AMC 的 R 值平均高出 8.8 个百分点.

原因分析:SRC-LDA 通过 must-link 中的局部特征词-情感词关联约束可以提高低频情感词的分配概率,从而提升关联于特征词的低频情感词的提取效果,如一些低频情感词“时尚”“霸气”“公道”和“合理”等.同时,一些中、低频情感词通过正、反义约束更好地匹配到了相对应的特征词,也提高了此类情感词的发现率,如与“便宜”对应的低频情感词“昂贵”“低廉”等.

3) 评价单元提取

评价单元提取的准确率和召回率可以考察模型的特征词-情感词关系聚合能力,即,具有评价关系的特征词

和情感词应尽量分配到同一主题.评价单元提取的 *P* 和 *R* 如图 14 所示,其中,横坐标为主题个数 *K*,纵坐标为准确率 *P* 或召回率 *R*.

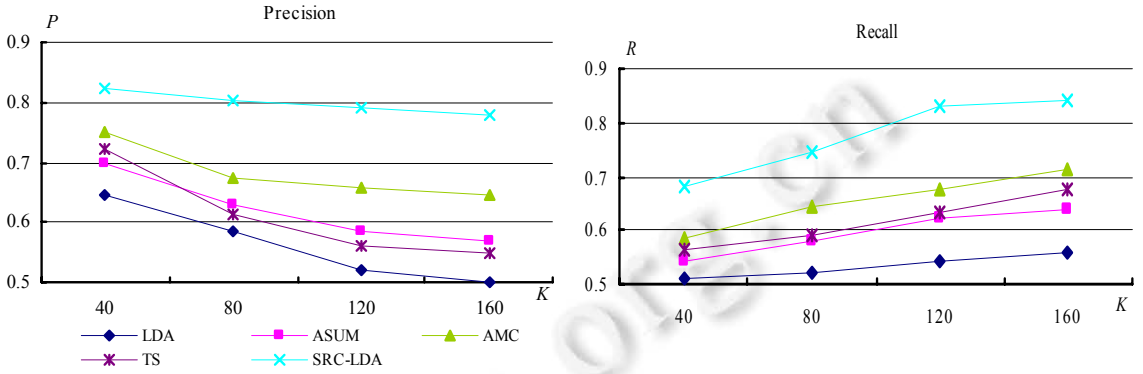


Fig.14 Precision and recall of appraisal expression extraction

图 14 评价单元提取的准确率和召回率

从图 14 可以看出:SRC-LDA 模型的准确率高于其他模型,且优势较明显,召回率的优势随着 *K* 的增加也更趋明显.关于准确率 *P* 的差异情况分析如下.

(1) 当 *K*=40 时,SRC-LDA 和 AMC 的 *P* 值相差最小,相差 7.2 个百分点;

(2) 当 *K*=160 时,SRC-LDA 和 AMC 的 *P* 值相差最大,相差 13.4 个百分点;

(3) SRC-LDA 比 AMC 的 *P* 值平均高出 11.7 个百分点.

原因分析:由于 LDA 倾向于发现全局特征词-情感词之间的关系,难以识别中、低频共现的局部特征词-情感词关系,所以准确率偏低;ASUM 和 TS 部分引入了主题词和情感词的关联,准确率均高于 LDA;AMC 没有从语义角度考查特征词-情感词的 must-link 约束,难以准确发现局部特征词-情感词之间的关联性;SRC-LDA 从语义角度获取特征词-情感词之间的 must-link 约束,增加了中、低频特征词-情感词之间关系发现的准确率.

关于召回率 *R* 的差异情况分析如下.

(1) 当 *K*=40 时,SRC-LDA 和 AMC 的 *R* 值相差最小,相差 9.4 个百分点;

(2) 当 *K*=120 时,SRC-LDA 和 AMC 的 *R* 值相差最大,相差 15.6 个百分点;

(3) SRC-LDA 比 AMC 的 *R* 值平均高出 12.1 个百分点.

原因分析:LDA 模型倾向于发现高频共现词语,这就导致了分配概率较高的词语在各主题下重复性较高,影响了中、低频特征词和情感词的识别;AUSM,TS 及 AMC 模型同样对应中、低频特征词和情感词具有不敏感性;对于 SRC-LDA 模型,一方面由于 cannot-link 约束提高了特征词和情感词的主题辨识度,减少了同一主题下特征词和情感词之间的错误匹配关系,另一方面,must-link 约束提高了主题下低频共现的特征词和情感词关系的识别率,使得能发现更多的局部特征词-情感词评价单元.例如,对于其他模型难以发现的一些低频共现词语关系,如(色彩,逼真)、(价格,公正)、(外观,圆润)和(颜色,饱满)等,SRC-LDA 模型利用特征词-情感词 must-link 可以识别这些评价单元.

4.4 SRC-LDA模型性能分析

为了对 SRC-LDA 模型性能进行更全面的分析,从以下两个方面考察不同约束知识对模型性能的影响:(1) 仅使用 must-link 约束,记为 M-SRC-LDA;(2) 仅使用 cannot-link 约束,记为 C-SRC-LDA.

上述两种情况和 SRC-LDA 的特征词、情感词和词语关联组合的准确率和召回率比较如图 15~图 17 所示,其中,横坐标为主题个数 *K*,纵坐标为准确率 *P* 或召回率 *R*.

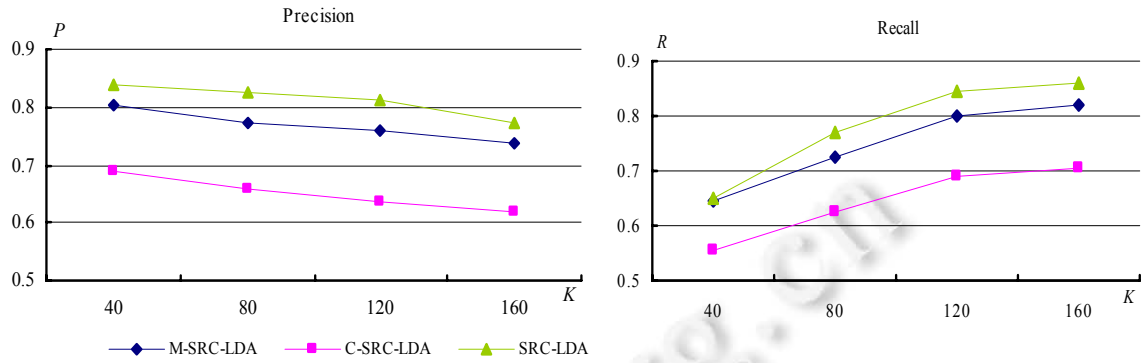


Fig.15 Precision and recall of aspect extraction

图 15 特征词提取的准确率和召回率

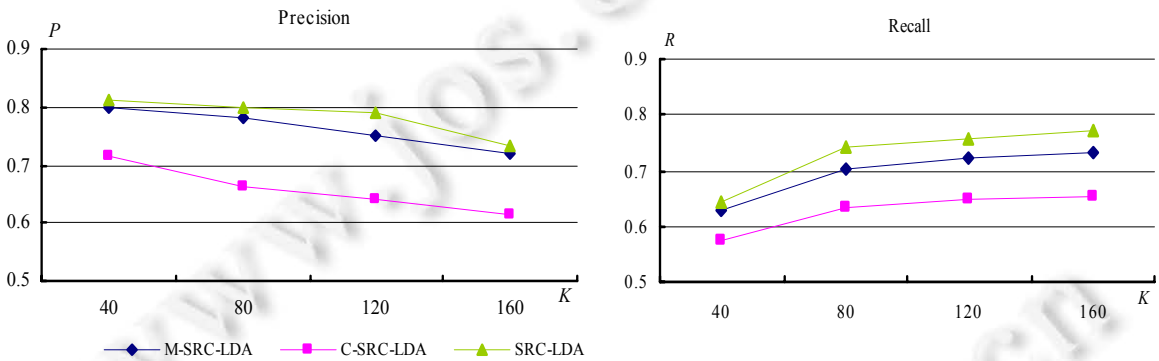


Fig.16 Precision and recall of opinion extraction

图 16 情感词提取的准确率和召回率

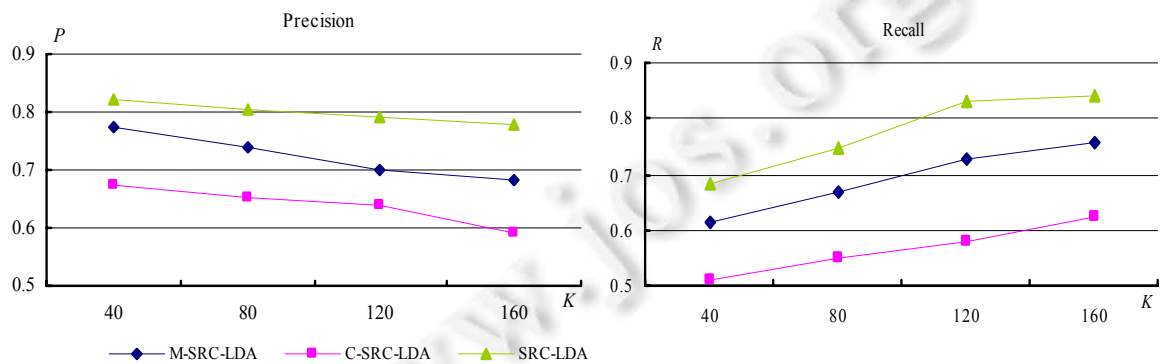


Fig.17 Precision and recall of appraisal expression extraction

图 17 评价单元提取的准确率和召回率

从图 15 可以看出:C-SRC-LDA 的准确率和召回率明显低于 M-SRC-LDA 和 SRC-LDA,M-SRC-LDA 和 SRC-LDA 之间相差不大.说明在特征词的提取中,仅使用 cannot-link 约束对于特征词的识别提升作用有限,而 must-link 约束对 LDA 模型产生了较大影响.原因在于 cannot-link 可以将不同的特征词尽量分配到不同主题,但一些低频特征词还是难以发现,而 must-link 利用同义关系可以更多地识别一些低频特征词,同时也将标准 LDA 中散布在各个主题中的全局特征词集中分配到少量主题,减少了这些高频特征词对中、低频局部特征词的分配干扰.

从图 16 可以看出:M-SRC-LDA 和 SRC-LDA 对情感词提取的准确率和召回率比较接近,因为利用特征词-情感词的 must-link 可以提高中、低频情感词的发现率,而使用 cannot-link 对情感词进行主题区分,可以增加主题对不同类情感词的识别度,但仅使用 cannot-link 不能有效提高中、低频情感词的识别率.

从图 17 可以看出:SRC-LDA 比 M-SRC-LDA,C-SRC-LDA 具有较明显的优势,说明同时使用 must-link 和 cannot-link 约束可以有效提升 LDA 对于特征词-情感词评价单元的识别性能;C-SRC-LDA 在召回率上明显低于其他模型,说明仅使用 cannot-link 约束虽然可以使不同类特征词和情感词尽量分配到不同主题,一定程度提高了特征词和情感词分配在同一主题的概率,但其关联性没有得到充分改善,还是难以发现一些低频的特征词-情感词关系;M-SRC-LDA 使用 must-link 约束可以增强局部特征词和情感词的关联性,同时可以发现更多的中、低频情感词,但缺少 cannot-link 约束容易导致不同类特征词和情感词在主题分配中的相互影响,尤其是来自全局特征词和全局情感词的干扰.

5 总结与展望

一方面,商品评论中存在很多低频的特征词和情感词,LDA 在主题分配中难以识别这类词语,使得低频特征和情感词的提取率不高;另一方面,商品评论文档中会同时出现多个不同特征的评价,LDA 很难辨别并将这些不相关特征分配到不同主题,造成特征词及其匹配的情感词没有实现较好的主题区分,难以有效提取特征和情感词.在中文商品评论中,特征词-特征词、特征词-情感词以及情感词-情感词之间蕴含着丰富的词语语义关系,可以利用这些语义知识来提高 LDA 对主题词语的识别度和区分度,改善标准 LDA 模型对一些低频特征词和情感词以及它们之间关系的提取率.

本文提出的 SRC-LDA 模型就是加入了 must-link 和 cannot-link 语义约束之后的 LDA 模型.一方面,通过 must-link 约束可以更多地发现低频的特征词和情感词,并将关联性强的特征词和情感词尽量分配到相同主题,提升中低频局部特征词和情感词的识别度,同时增加了词语间的主题聚合度;另一方面,通过 cannot-link 约束可以更多地发现评无相关的特征词和情感词,并将无关联的特征词和情感词尽量分配到不同主题,提升特征词和情感词的区分度.

实验结果表明:SRC-LDA 模型改善了特征词和情感词的主题内聚度,改进了特征词和情感词的主题区分度,从而提高了特征词、情感词和评价单元的提取率.对于特征词提取,SRC-LDA 模型比 AMC[27]的准确率平均高出 11.7 个百分点、召回率平均高出 11 个百分点;对于情感词提取,SRC-LDA 模型比 AMC 的准确率平均高出 10.1 个百分点、召回率平均高出 8.8 个百分点;对于评价单元提取,SRC-LDA 模型比 AMC 的准确率平均高出 11.7 个百分点、召回率平均高出 12.1 个百分点.

下一步工作希望继续挖掘中文商品评论中的语义知识来影响主题模型对于主题词的提取,更多发现符合语义要求的特征词和情感词.

References:

[1] Liu B. Sentiment Analysis and Opinion Mining. California: Morgan & Claypool Publishers, 2012. [doi: 10.2200/S00416ED1V01Y201204HLT016]

[2] Hu MQ, Liu B. Mining opinion features in customer reviews. In: Proc. of the 19th National Conf. on Artificial Intelligence (AAAI 2004). AAAI Press, 2004. 755−760.

[3] Popescu AM, Etzioni O. Extracting product features and opinions from reviews. In: Proc. of the Human Language Technology Conf. and the Conf. on Empirical Methods in Natural Language Processing (HLT/EMNLP 2005). Vancouver: ACL Press, 2005. 339−346. [doi: 10.1007/978-1-84628-754-1_2]

[4] Jakob N, Gurevych I. Extracting opinion targets in a single and cross-domain setting with conditional random fields. In: Proc. of the Conf. on Empirical Methods in Natural Language Processing (EMNLP 2010). Cambridge: ACL Press, 2010. 1035−1045.

[5] Jin W, Ho HH, Srihari RK. OpinionMiner: A novel machine learning system for Web opinion mining and extraction. In: Proc. of the 15th ACM Int'l Conf. on Knowledge Discovery and Data Mining (SIGKDD 2009). Paris: ACM Press, 2009. 1195−1204. [doi: 10.1007/978-1-84628-754-1_2]

[6] Su Q, Xu XY, Guo HL, Guo ZL, WuX, Zhang XX, Swen B, Su Z. Hidden sentiment association in Chinese web opinion mining. In: Proc. of the 17th Int'l Conf. on World Wide Web (WWW 2008). Beijing: ACM Press, 2008. 959−968. [doi: 10.1145/1367497.1367627]

[7] Wang RY, Ju JP, Li SS, Zhou GD. Feature engineering for CRFs based opinion target extraction. Journal of Chinese Information Processing, 2012,26(2):56−61 (in Chinese with English abstract).

[8] Liu HY, Zhao YY, Qin B, Liu T. Comment target extraction and sentiment classification. Journal of Chinese Information Processing, 2010,24(1):84−88 (in Chinese with English abstract).

[9] Wu YB, Zhang Q, Huang XJ, Wu LD. Phrase dependency parsing for opinion mining. In: Proc. of the 47th Annual Meeting of the Association for Computational Linguistics (ACL 2009). Singapore: ACL Press, 2009. 1553−1541.

[10] Zhao YY, Qin B, Che WX, Liu T. Appraisal expression recognition based on syntactic path. Ruan Jian Xue Bao/Journal of Software, 2011,22(5):887−898 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/3767.htm [doi: 10.3724.SP.J.1001.2011.03767]

[11] Qiu G, Liu B, Bu JJ, Chen C. Opinion word expansion and target extraction through double propagation. Computational Linguistics, 2011,37(1):9−27. [doi: 10.1162/coli_a_00034]

[12] Yao TF, Nie QY, Li JC, Li LL, Lou DC, Chen K, Fu Y. An opinion mining system for Chinese automobile reviews. In: Proc. of the Frontiers of Chinese Information Processing. Beijing: Tsinghua University Press, 2006. 260−281.

[13] Poria S, Cambria E, Ku LW, Gui C, Gelbukh A. A rule-based approach to aspect extraction from product reviews. In: Proc. of the 2nd Workshop on Natural Language Processing for Social Media (SocialNLP 2014). Dublin: ACM Press, 2014. 28−37. [doi: 10.3115/v1/W14-5905]

[14] Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. Journal of Machine Learning Research, 2003,3(3):993−1022.

[15] Titov I, McDonald RT. Modeling online reviews with multi-grain topic models. In: Proc. of the 17th Int'l Conf. on World Wide Web (WWW 2008). Beijing: ACM Press, 2008. 111−120. [doi: 10.1145/1367497.1367513]

[16] Andrzejewski D, Zhu X, Craven M. Incorporating domain knowledge into topic modeling via Dirichlet forest priors. In: Proc. of the 26th Annual Int'l Conf. on Machine Learning (ICML 2009). Montreal: ACM Press, 2009. 25−32. [doi: 10.1145/1553374.1553378]

[17] Zhai ZW, Liu B, Xu H, Jia PF. Constrained LDA for grouping product features in opinion mining. In: Proc. of the Pacific-Asia Conf. on Knowledge Discovery and Data Mining (PAKDD 2011). Shenzhen: Springer-Verlag Press, 2011. 448−459. [doi: 10.1007/978-3- 642-20841-6_37]

[18] Chen ZY, Mukherjee A, Liu B, Hsu M, Castellanos M, Ghosh R. Exploiting domain knowledge in aspect extraction. In: Proc. of the Conf. on Empirical Methods on Natural Language Processing (EMNLP 2013). Seattle: ACL Press, 2013. 1655−1667.

[19] Bagheri A, Saraee M, Jong FD. ADM-LDA: An aspect detection model based on topic modeling using the structure of review sentences. Journal of Information Science, 2014,40(5):621−636. [doi: 10.1177/0165551514538744]

[20] Ma BZ, Yan ZJ. Product features extraction of online reviews based on LDA model. Computer Integrated Manufacturing Systems, 2014,20(1):96−103 (in Chinese with English abstract).

[21] Chen ZY, Mukherjee A, Liu B. Aspect extraction with automated prior knowledge learning. In: Proc. of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014). Baltimore: ACL Press, 2014. 347−358. [doi: 10.3115/v1/P14-1033]

[22] Lin C, He Y. Joint sentiment/topic model for sentiment analysis. In: Proc. of the 18th ACM Conf. on Information and Knowledge Management (CIKM 2009). Hongkong: ACM Press, 2009. 375−384. [doi: 10.1145/1645953.1646003]

[23] Lu B, Ott M, Cardie C, Tsou BK. Multi-Aspect sentiment analysis with topic models. In: Proc. of the 11th IEEE Int'l Conf. on Data Mining (ICDM 2011). Vancouver: IEEE Press, 2011. 81−88. [doi: 10.1109/ICDMW.2011.125]

[24] Jo Y, Oh AH. Aspect and sentiment unification model for online review analysis. In: Proc. of the 4th ACM Int'l Conf. on Web Search and Data Mining (ICWS 2011). Washington: IEEE Press, 2011. 815−824. [doi: 10.1145/1935826.1935932]

[25] Moghaddam S, Ester M. ILDA: Interdependent LDA model for learning latent aspects and their ratings from online product reviews. In: Proc. of the 34th Int'l Conf. on Research and Development in Information Retrieval (SIGIR 2011). Beijing: ACM Press, 2011. 665−674. [doi: 10.1145/2009916.2010006]

[26] Sun Y, Zhou XG, Fu W. Unsupervised topic and sentiment unification model for sentiment analysis. Acta Scientiarum Naturalium Universitatis Pekinensis, 2013,49(1):102−108 (in Chinese with English abstract).

[27] Chen ZY, Liu B. Mining topics in documents: Standing on the shoulders of big data. In: Proc. of the 20th Int'l Conf on Knowledge Discovery and Data Mining (SIGKDD 2014). New York: ACM Press, 2014. 1116−1125. [doi: 10.1145/2623330.2623622]

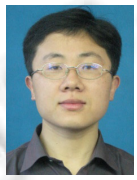
[28] Dermouche M, Kouas L, Velcin J, Loudcher S. A joint model for topic-sentiment modeling from text. In: Proc. of the Symp. on Applied Computing (SAC 2015). Salamanca: ACM Press, 2015. 819−824. [doi: 10.1145/2623330.2623622]

[29] Ouyang JH, Liu YH, Li XM, Zhou XT. Multi-Grain sentiment/topic model based on LDA. Acta Electronica Sinica, 2015,43(9): 1875−1880 (in Chinese with English abstract).
[30] Li FT, Huang ML, Zhu XY. Sentiment analysis with global topics and local dependency. In: Proc. of the 24th Conf. on Artificial Intelligence (AAAI 2010). Atlanta: AAAI Press, 2010. 1371−1376.
[31] Zhao WX, Jiang J, Yan HF, Li XM. Jointly modeling aspects and opinions with a MaxEnt-LDA hybrid. In: Proc. of the 2010 Conf. on Empirical Methods in Natural Language Processing (EMNLP 2010). Cambridge: ACL Press, 2010. 56−65.
[32] Mukherjee A, Liu B. Aspect extraction through semi-supervised modeling. In: Proc. of the 50th Annual Meeting of the Association for Computational Linguistics (ACL 2012). Jeju Island: ACL Press, 2012. 339−348.
[33] Ma C, Wang M, Chen X. Topic and sentiment unification maximum entropy model for online review analysis. In: Proc. of the 24th Int'l Conf. on World Wide Web Companion (WWW 2015). Florence: ACM Press, 2015. 649−654. [doi: 10.1145/2740908.2741704]
[34] Li FT, Wang S, Liu SH, Zhang M. SUIT: A supervised user-item based topic model for sentiment analysis. In: Proc. of the 28th Conf. on Artificial Intelligence (AAAI 2014). Quebec: AAAI Press, 2014. 1636−1642.
[35] Heyrani-Nobari G, Chua TS. User intent identification from online discussions using a joint aspect-action topic model. In: Proc. of the 28th Conf. on Artificial Intelligence (AAAI 2014). Quebec: AAAI Press, 2014. 1221−1227.
[36] Yang ZH, Kotov A, Mohan A, Lu SY. Parametric and non-parametric user-aware sentiment topic models. In: Proc. of the 38th Int'l Conf. on Research and Development in Information Retrieval (SIGIR 2015). Santiago: ACM Press, 2015. 413−422. [doi: 10.1145/2766462.2767758]
[37] Che WX, Li ZH, Liu T. LTP: A Chinese language technology platform. In: Proc. of the 23rd Int'l Conf. on Computational Linguistics: Demonstrations (COLING 2010). Beijing: ACM Press, 2010. 3−16.

附中文参考文献:

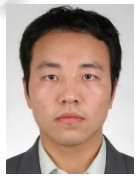
[7] 王荣洋,鞠久朋,李寿山,周国栋.基于 CRFs 的评价对象抽取特征研究.中文信息学报,2012,26(2):56−61.
[8] 刘鸿宇,赵妍妍,秦兵,刘挺.评价对象抽取及其倾向性分析.中文信息学报,2010,24(1):84−88.
[10] 赵妍妍,秦兵,车万翔,刘挺.基于句法路径的情感评价单元识别.软件学报,2011,22(5):887−898. http://www.jos.org.cn/1000-9825/3767.htm [doi: 10.3724.SP.J.1001.2011.03767]
[12] 姚天昉,聂青阳,李建超,李林琳,娄德成,陈珂,付宇.一个用于汉语汽车评论的意见挖掘系统.见:中文信息处理前沿进展会议论文集.北京:清华大学出版社,2006.260−281.
[20] 马柏樟,颜志军.基于潜在狄利特雷分布模型的网络评论产品特征抽取方法.计算机集成制造系统,2014,20(1):96−103.
[26] 孙艳,周学广,付伟.基于主题情感混合模型的无监督文本情感分析.北京大学学报(自然科学版),2013,49(1):102−108.
[29] 欧阳继红,刘燕辉,李熙铭,周晓堂.基于 LDA 的多粒度主题情感混合模型.电子学报,2015,43(9):1875−1880.

彭云(1972－),男,江西宜春人,博士生,副教授,CCF 专业会员,主要研究领域为情感分析,数据挖掘,自然语言处理.

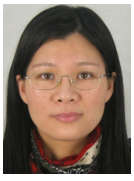


刘德喜(1975－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为 Web 数据管理,信息检索,自然语言处理.

万常选(1962－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为 Web 数据管理,情感分析,信息检索,数据挖掘.



刘喜平(1981－),男,博士,副教授,CCF 专业会员,主要研究领域为信息检索,数据挖掘.



江腾蛟(1976－),女,博士,讲师,主要研究领域为情感分析,Web 数据管理.

廖国琼(1969－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为数据库,数据挖掘,社会网络.