软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2018,29(10):2899−2914 [doi: 10.13328/j.cnki.jos.005548]

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

融合文本概念化与网络表示的观点检索*

廖祥文[1,2], 刘德元[1,2], 桂 林[1,2], 程学旗[3], 陈国龙[1,2]

[1](福州大学 数学与计算机科学学院,福建 福州 350116)

[2](福建省网络计算与智能信息处理重点实验室(福州大学),福建 福州 350116)

[3](网络数据科学与技术重点实验室(中国科学院),北京 100190)

通讯作者: 桂林, guilin.nlp@gmail.com

摘 要: 观点检索是自然语言处理领域中的一个热点研究课题.现有的观点检索模型在检索过程中往往无法根据上下文将词汇进行知识、概念层面的抽象,在语义层面忽略词汇之间的语义联系,观点层面缺乏观点泛化能力.因此,提出一种融合文本概念化与网络表示的观点检索方法.该方法首先利用知识图谱分别将用户查询和文本概念化到正确的概念空间,并利用网络表示将知识图谱中的词汇节点表示成低维向量,然后根据词向量推出查询和文本的向量,并用余弦公式计算用户查询与文本的相关度,接着引入基于统计机器学习的分类方法挖掘文本的观点.最后,利用概念空间、网络表示空间以及观点分析结果构建特征,并服务于观点检索模型.相关实验结果表明,所提出的检索模型可以有效提高多种检索模型的观点检索性能.其中,基于统一相关模型的观点检索方法在两个实验数据集上相比于基准方法,在 MAP 评价指标上分别提升了 6.1%和 9.3%,基于排序学习的观点检索方法在两个实验数据集上相比于基准方法,在 MAP 评价指标上分别提升了 2.3%和 14.6%.

关键词: 信息检索;观点检索;知识图谱;文本概念化;网络表示

中图法分类号: TP311

中文引用格式: 廖祥文,刘德元,桂林,程学旗,陈国龙.融合文本概念化与网络表示的观点检索.软件学报,2018,29(10):2899−2914. http://www.jos.org.cn/1000-9825/5548.htm

英文引用格式: Liao XW, Liu DY, Gui L, Cheng XQ, Chen GL. Opinion retrieval method combining text conceptualization and network embedding. Ruan Jian Xue Bao/Journal of Software, 2018,29(10):2899−2914 (in Chinese). http://www.jos.org.cn/1000-9825/5548.htm

Opinion Retrieval Method Combining Text Conceptualization and Network Embedding

LIAO Xiang-Wen[1,2], LIU De-Yuan[1,2], GUI Lin[1,2], CHENG Xue-Qi[3], CHEN Guo-Long[1,2]

[1](College of Mathematics and Computer Science, Fuzhou University, Fuzhou 350116, China)

[2](Fujian Provincial Key Laboratory of Networking Computing and Intelligent Information Processing (Fuzhou University), Fuzhou 350116, China)

[3](Key Laboratory of Network Data Science and Technology (The Chinese Academy of Sciences), Beijing 100190, China)

* 基金项目: 国家自然科学基金(61772135, U1605251); 中国科学院网络数据科学与技术重点实验室开放基金(CASNDST 201708, CASNDST201606); 可信分布式计算与服务教育部重点实验室主任基金(2017KF01); 福建省自然科学基金(2017J01755); 赛尔网络下一代互联网技术创新项目(NGII20160501)

Foundation item: National Natural Science Foundation of China (61772135, U1605251); Open Project of Key Laboratory of Network Data Science & Technology of the Chinese Academy of Sciences (CASNDST201708, CASNDST201606); Director's Project Fund of Key Laboratory of Trustworthy Distributed Computing and Service (BUPT), Ministry of Education (2017KF01); Natural Science Foundation of Fujian Province of China (2017J01755); CERNET Innovation Project (NGII20160501)

本文由"本体工程与知识图谱"专题特约编辑漆桂林教授推荐.

收稿时间: 2017-07-20; 修改时间: 2017-11-08; 采用时间: 2018-01-24; jos 在线出版时间: 2018-02-08

CNKI 网络优先出版: 2018-02-08 11:55:38, http://kns.cnki.net/kcms/detail/11.2560.TP.20180208.1155.004.html

Abstract: Opinion retrieval is a hot topic in the research of natural language processing. Most existing approaches in text opinion retrieval can not extract knowledge and concept from context. They also lack opinion generalization ability and overlook the semantic relations between words. This paper proposes an opinion retrieval method based on knowledge graph conceptualization and network embedding. First, conceptual knowledge graph is used to conceptualize the queries and texts into the correct conceptual space while the nodes in the knowledge graph are embedded into low dimensional vectors space by network embedding technology. Then, the similarity between queries and texts is calculated based on embedding vectors. According to the similarity score, the opinion scores of texts can be captured based on statistical machine learning methods. Finally, the concept space, knowledge representation space, and opinion mining result serve opinion retrieval models. The experiment shows that the retrieval model proposed in this paper can effectively improve the retrieval performance of multiple retrieval models. Compared with referenced method based on unified opinion, the proposed approach improves the MAP scores by 6.1% and 9.3%, respectively. Compared with referenced method based on learning to rank, proposed approach improves the MAP scores by 2.3% and 14.6%, respectively.

Key words: information retrieval; opinion retrieval; knowledge graph; text conceptualization; network embedding

随着互联网的迅猛发展,网络中涌现了大量的论坛、博客等社交媒体,吸引大量用户在这些社交媒体上分享他们关于政治、产品、公司、事件的观点.观点检索旨在从社交媒体等文档集中检索出与查询主题相关并且表达用户观点(赞同或反对)的文档,是自然语言处理领域里的一项重要课题[1,2].

国际文本检索会议(The Text Retrieval Conf.,简称 TREC)在 2006 年开始引入博客观点检索的评测,之后涌现了大量的观点检索方面的研究[3−7].早期研究的观点检索是两阶段模型[3,4]:首先,利用传统的信息检索模型获得与查询相关的候选相关文档;然后,将候选相关文档根据观点得分进行重排序.之后出现了将主题相关度与观点结合起来的统一相关模型(unified relevance model)[5−7].该模型借助当前信息检索和观点挖掘领域的最新模型,直接挖掘描述查询的观点对文档进行排序.后来出现了排序学习模型(learning to rank,简称 L2R)[8,9],利用提取的特征和机器学习的方法对推文进行倾向性检索.

但上述排序学习模型往往产生较为稀疏的特征空间,统一相关模型在检索的过程中泛化能力有一定缺陷,这种缺陷主要体现在 3 个方面.

第一,在知识层面,上述模型往往无法根据上下文将词汇进行知识、概念层面的抽象.例如:

例:Ios5 update gets android like notification bar!? apple bowed to google!

译:Ios5 更新得到类似 Android 的通知栏!苹果向谷歌低头了!

上述文本提到“apple\苹果”,现有模型无法识别其是指苹果公司还是苹果这种水果.因此,上述统一相关模型缺乏知识的泛化.

第二,在语义层面,上述检索模型处理查询未登录词并且由此引出的数据稀疏性问题的能力有限,查询扩展技术[3,10]虽然扩展了原始查询的语义信息,但整体检索仍是词匹配的过程,无法计算词汇之间的语义相似度,缺乏计算的泛化.

第三,在观点层面,上述基于词典的观点挖掘方法受限于观点词典的覆盖率限制,无法处理未登录的观点词汇,缺乏观点泛化能力.

针对上述问题,本文提出了融合文本概念化与网络表示的观点检索模型:首先引入概念知识图谱,通过有效分析查询和文本的概念空间,判断对应多个概念的实体在具体上下文中的概念,以此来实现概念级别的推理,提高检索模型知识泛化的能力;其次,通过基于网络节点的网络表示学习,有效地利用知识图谱中的结构化信息,学习捕获词汇之间的语义信息,把词汇节点投射到低维的语义空间中,这使得在传统的词匹配中,词汇之间由于特征稀疏所引起的语义相似度计算困难现象通过低维空间中向量计算得以一定程度的改善,能够提高语义计算的泛化能力;最后,通过引入朴素贝叶斯支持向量机(Naïve Bayes support vector machines,简称 NBSVMs[11])和卷积神经网络(convolutional neutral network,简称 CNN[12])方法挖掘文本的观点,摆脱了基于观点词典方法泛化能力有限的制约,进一步提高了观点检索的性能.基于上述 3 种特征表示与观点建模方法,本文进一步将 3 类特征应用于统一相关模型以及基于排序学习的观点检索模型.相关实验结果表明:本文提出的 3 类特征表示与观点建模方法可以有效提高观点检索的性能,并且不论具体的应用场景中是否提供了有标注训练集,本文提出的

方法均能有效提高现有的观点检索精度,是一种通用性很好的观点检索方法.

本文第 1 节介绍已有研究的相关工作.第 2 节首先介绍本文的问题描述和方法概述,然后详细描述主要模块的细节.第 3 节为实验设置与结果分析,通过与基准工作对比来验证本文方法的有效性.第 4 节为总结.

1 相关工作

目前,观点检索研究主要包括两方面的内容.一方面是如何对信息检索中的文档与查询词进行特征表示,另一方面是如何构建检索模型度量查询词与待检索文档之间的观点得分.针对上述两方面内容,本节将分别介绍目前国内外相关的研究工作.

1.1 特征表示

传统的信息检索模型主要将查询词与待检索文档映射到某个高维向量空间进行相似度的计算,以度量不同文本之间的相似度,并根据相似度结果返回文档的排名用于信息检索,例如 BM25(基于 Okapi BM25[13])和 VSM(基于 vector space model[14]).但是此类方法均是基于词袋模型的特征表示,无法对词汇进行语义、概念层面的分析.并且由于文本和用户查询通常比较简短,产生的特征空间比较稀疏,导致文本间的语义相似度计算困难,泛化能力有限.因此,有研究者使用知识图谱、文本概念化和网络表示技术对文本进行语义理解,构建文本的语义特征表示,应用于检索任务.

近年来,随着各种知识获取和知识图谱构建技术的逐渐完善[15,16],关于知识图谱的应用研究引起了很多研究者的兴趣.Dalton 等人[17]利用实体的相关特征和实体与知识库的连接(包括结构化的属性和文本)来丰富原始查询.Xiong 等人[18]提出利用 freebase 获取与查询相关的实体,然后利用非监督或者监督的方法得到最终的扩展词.Wang[19]提出文本概念化模型,借助知识图谱对文本进行解析和推理,进而将其映射到知识图谱中的一组概念上,在文本分类任务上取得 90%+的准确率.王仲远[20]提出借助知识图谱为文本构建统一的候选词关系图,并使用随机漫步(random walk)的方法推导出最优的分词、词性和词的概念,提高知识泛化准确率.另有学者利用知识图谱提高概念漂移检测[21]和问答系统[22,23]的性能.

网络表示学习是面向知识图谱中的实体和关系的表示学习[24],在多关系知识图谱表示中,Bordes 等人提出的 TransE 模型[25]引起了广泛关注和扩展,将知识图谱中的关系 r 解释为头实体 h 到尾实体 t 的翻译操作,认为向量 $h+r$ 应该靠近向量 t.但是 TransE 无法处理复杂关系(如一个头实体对应多个尾实体、多对一、多对多),为此,Wang 等人提出了 TransH 模型[26],让实体在不同的关系下有不同的表示,有效解决了 TransE 的缺点.针对 TransH 仍存在将实体和关系映射在同一语义空间这一缺陷,Lin 等人提出了 TransR 模型[27],将实体和关系分别建模在实体空间和关系空间,并在关系空间执行翻译.TransR 较 TransE 和 TransH 有很多改进,但仍存在参数过多、计算复杂度高的缺点,为此,Ji 等人提出了 TransD 模型[28],利用两个投影向量构建投影矩阵,解决了 TransR 模型计算复杂度高的问题.随后,Ji 又提出了 TranSparse 模型[29],针对不同复杂度的关系,使用不同稀疏程度的矩阵进行表征,以防止对关系的过拟合或者欠拟合现象.He 等人认为,实体和关系可能存在不同的语义,以往的模型忽略了语义的不确定性,为此提出了 KG2E 模型[30],使用高斯 embedding 进行知识表示学习.

在单一关系知识图谱表示领域,Ahmed[31]提出了 GF,将信息网络表示成关联矩阵,通过矩阵分解将节点表示到低维稠密的向量空间.Perozzi[8]提出了 DeepWalk,将节点视为单词,将在网络上随机游走的路径视为句子,获得的数据直接作为 word2vec 算法的输入以训练节点的向量表示.Tang[9]提出了 LINE,直接针对网络的一阶相似度和二阶相似度进行建模,有效保留了网络的结构信息.Jacob[32]提出了 LSHM 模型,在训练节点的向量表示时,同时考虑了分类函数对已知节点标签的分类能力.

1.2 检索模型

目前的观点检索模型大体上可以分为 3 类.

第 1 类观点检索方法是两阶段模型:第 1 阶段使用传统信息检索方法得到主题相关文档,第 2 阶段对主题相关文档计算观点得分.例如,Zhang 等人[3]首先利用传统信息检索模型和查询扩展技术找出主题相关的文档,

接着,用支持向量机(SVMs)分类器对主题相关文档进行观点分类并重排序.Santos 等人[33]首先利用两种现有方法找出观点语句,接着,将查询与观点语句的邻近关系融入到 DFR(divergence from randomness)邻近关系模型中,最终得到文档的观点检索评分.Wang 等人[4]把重点放在观点分类方面,通过整合推文、Hashtag 间的共现关系等特征,采用 3 种图模型的分类算法进行观点分类.

第 2 类方法是统一相关检索模型,该模型直接挖掘描述查询的观点得分,对文档排序,相对于两阶段模型,具有理论上易解释、对信息需求表达更直接、有效的优点.例如,Eguchi[34]提出一种概率生成模型框架下的观点检索模型,通过考虑查询依赖的观点得分,将主题相关模型与观点得分结合起来,进而计算文档的排序得分.Zhang[5]提出一个基于词典的生成模型,通过二次组合方式(quadratic combination)将主题相关得分与观点评分结合.但该模型假设观点词是均匀分布的.Liao[6]考虑了观点词所含观点信息的差异性,首先基于异质图计算观点词权重,然后将其融入生成模型.文献[10]则利用外源知识和机器学习的方法扩展用户的查询词并融入生成模型. Huang[7]通过查询相关与查询无关的混合倾向性扩展,将主题检索与倾向性分类的两阶段方法转换成一个统一的观点检索进程.但大部分统一观点检索模型忽略了对用户查询和文本的语义分析,存在仅考虑了词语的表面匹配、不能处理同义词和一词多义等问题.

第 3 类方法是排序学习模型(learning to rank,简称 L2R).Luo[35]利用文档特征、博主特征和主观性特征,采用排序学习模型对推文进行观点检索.Kim[36]进一步利用了博主特征和标签特征的主观性信息来描述文档的主观倾向.一般而言,使用排序学习算法进行信息检索往往可以取得较高的精度,但由于其需要大量的人工标注数据构建训练集,因此这一方法的应用场景相对于前两种方法而言较为有限;并且该模型针对不同的查询,其相关文档数量的差异会对学习的效果评价造成偏置.

本文从观点检索目前相关研究的 3 种局限性出发,希望可以借助知识库资源以及相关机器学习方法,并结合情感模型,进行更为有效的特征表示.同时,希望学习获得的特征可以普适性地提高不同的信息检索模型性能,以达到面向不同应用场景的通用观点检索的目的.

2 融合文本概念化与网络表示的观点检索

本文提出一种融合文本概念化与网络表示的观点检索:首先,利用知识图谱分别将用户查询和文本集概念化到概念集合上,同时,利用网络表示技术将知识图谱中的节点表示成低维向量;然后,通过逐点的向量相加并取均值的方式推出文本向量和查询向量,并使用余弦公式计算查询向量和文本向量的相关度得分,接着引入 NBSVMs 和 CNN 两种分类方法计算文本观点得分;最后,将文本概念化结果、网络表示结果、观点得分结果作为特征,进行观点检索模型的设计.具体而言,本文在上述特征的基础上,分别使用了不需要训练语料的基于统一相关模型的观点检索方法以及需要训练语料的基于排序学习模型的观点检索方法进行观点检索,验证了本文提出的相关方法的有效性.

2.1 问题描述

为了方便研究,本节将基于统一相关模型的观点检索研究问题形式化地描述为:给定一个查询 q、观点词典 $T=\{t_i,i=1,2,\ldots,M\}$(其中,t_i 表示观点词及其评分,M 表示观点词典的大小)、待检索的文档集合 $D=\{d_i,i=1,2,\ldots,N\}$(其中,d_i 表示文档 i 的文本内容,N 表示文档集的大小)以及知识图谱 $G=(V,E)$(其中,V 表示知识图谱中的节点集合,包括实体集和概念集;E 表示知识图谱中的边的集合,每条边表示一个实体-概念对(entity-concept pair)),计算每个待检索文档 d_i 与查询 q 的主题相关度得分 $I_{rel}=(d,q)$和 d_i 的观点得分 $I_{opn}=(d,q,T)$;根据检索模型将相关度得分和观点得分二次组合得到最终的相关观点得分 $Rank(d)=(d,q,G,T)$,并根据相关观点评分从高到低排序.

2.2 基于知识图谱的文本概念化

文本概念化的目的是借助概念知识图谱推理出文本中每个实体的概念分布,即,将实体按照其上下文语境映射到正确的概念集合上[19](bags-of-concepts,简称 BOC).

例:Ios5 update gets android like notification bar!? apple bowed to google!

译:Ios5 更新得到类似 Android 的通知栏!苹果向谷歌低头了!

在上述文本中,通过知识图谱 Probase[37],机器可获悉“apple\苹果”这个实体有“fruit\水果”和“company\公司”等概念,“google\谷歌”这个实体有“company\公司”等概念.当“apple\苹果”与“google\谷歌”同时出现在文本中时,通过贝叶斯公式可以分析出该文本中的“apple\苹果”有较高的概率属于“company\公司”这一概念.

给定文档集合 $D=\{d_i,i=1,2,\ldots,N\}$,本文利用 Probase 推理每篇文档的概念集合.文档的相关概念最终表示为一个概念集合 $d_i=(\langle c_1,w_1\rangle,\ldots,\langle c_j,w_j\rangle,\ldots,\langle c_k,w_k\rangle),i=1,2,\ldots,N,j=1,2,\ldots,k$,其中,$w_j$ 表示概念 c_j 属于该文档的权重,反映了概念 c_j 对该文档的解释能力.概念化过程分为两部分:实体识别、概念推理.

2.2.1 基于逆向最大匹配的实体识别

为了获得文本的概念集合,首先需要识别文本中的实体,以便通过实体推理概念.对于多词表达的实体,本文仅考虑长度最大的一项,实体之间不相互包含.例如“apple inc\苹果公司”可能有两种实体识别结果:“apple\苹果”“inc\公司”或者“apple inc\苹果公司”,因为三者都在词典中,但本文仅考虑“apple inc\苹果公司”这一实体.因此,采用基于词典的逆向最大匹配算法来识别每篇文档中的实体,并选用知识图谱 Probase 中的所有实体(约 1 200 万个实体)作为匹配词典.匹配过程中,采用波特提取器(http://tartarus.org/~martin/PorterStemmer/)对文档和词典分别做词干提取处理.具体算法描述如下.

算法. 基于逆向最大匹配的实体识别算法.

输入:文档集合、实体词典;

输出:每篇文档的实体集合.

初始化:对实体词典每个词项做词干提取处理.设词典中实体最大长度(包含词汇个数)为 max*Len*,设输出实体集合 *entitySet* 为空.对每篇文档进行如下处理.

Step 1:对文档词汇做词干提取处理,得到文本 $s=s_1s_2\ldots s_n$.

Step 2:计算 s 包含词汇个数,设为 n:如果 n 等于 0,转 Step 7;如果 $n<\max Len$,则 $\max Len=n$.

Step 3:取出 $str=s_{n-\max Len}\ldots s_n$ 作为候选实体.

Step 4:查看 *str* 是否在词典中:若是,则转 Step 5;否则,转 Step 6.

Step 5:将 *str* 加入 *entitySet* 中,$s=s-str$,转 Step 2.

Step 6:如果 *str* 长度等于 1,$s=s-str$,转 Step 2;否则,将 *str* 最左边的一个词去掉,转 Step 4.

Step 7:输出 *entitySet*,结束.

2.2.2 基于朴素贝叶斯模型的概念推理

给定文档的实体集合 $E=\{e_i,i=1,2,\ldots,M\}$,概念生成的目的是利用 Probase 中的实体-概念对(instance-concept pairs)推理出最能描述该实体集合的概念集合.为评估概念对文档的表示能力,采用朴素贝叶斯模型进行评估:

$$P(c_k\mid E)=\frac{P(E\mid c_k)P(c_k)}{P(E)}\propto P(c_k)\prod_{i=1}^{M}P(e_i\mid c_k) \tag{1}$$

通过贝叶斯公式计算每个概念的后验概率,获得高后验概率值的概念显然就是最能代表给定实体集合的概念.同时,把后验概率值作为这个概念以表达该文档的解释能力,即为该概念的权重.

在公式(1)中,给定概念,得到实体的概率公式为

$$P(e_i\mid c_k)=\frac{n(e_i,c_k)}{n(c_k)} \tag{2}$$

其中,$n(e_i,c_k)$表示 e_i 和 c_k 的共现次数,$n(c_k)$表示 c_k 出现的次数.这两个值都可以从 Probase 中直接或计算得到.两个文本例子及它们经概念化后的概念集合见表 1.

Table 1 Samples of text conceptualization

表 1 文本概念化样例

文本	概念
在美国音乐奖上赞成克里斯·布朗赢得最佳男艺人奖.	(1) “克里斯·布朗”:歌手,明星,说唱歌手,畅销男歌手,黑人男歌手… (2) “最佳男歌手”:奖项 (3) “美国音乐奖”:颁奖仪式,好莱坞事件,电视节目,电视音乐节目…

Table 1 Samples of text conceptualization (Continued)
表 1 文本概念化样例(续)

文本	概念
iOS5 更新得到类似 Android 的通知栏! 苹果向谷歌低头了!	(1) “iOS5”:系统,先进技术,苹果设备… (2) “Android”:系统,手机操作系统,移动设备,智能手机… (3) “通知栏”:用户界面元素,安卓特有元素 (4) “苹果”:公司,品牌,科技公司,科技巨头… (5) “谷歌”:公司,科技公司,科技巨头,搜索引擎…

2.3 信息网络表示(information network embedding)

经过文本概念化获得文本和查询的概念集合.这种模型下的主题相似度计算仍是以匹配为主,无法处理概念之间的语义相似度.例如,“Company\公司”和“Organization\组织”,“Company\公司”是“Organization\组织”的一个子概念,它们具有高度语义相似性,但这种信息在概念匹配时就会丢失.网络表示(network embedding,简称NE)可以将网络节点表示成低维向量,同时保留节点之间的这种相似性信息.这样,在低维空间里可以高效计算概念之间的语义相似度.因此,用网络表示改善概念匹配的缺点,提高模型的计算泛化能力.

图 1 所示为 Probase 网络的一个子图例子.两个节点的边表示的是网络的局部特性,边的权值通常表明这两个节点的关联程度.例如,概念“Company\公司”和“Organization\组织”有边相连,在语义上“Company\公司”是“Organization\组织”的子概念,因此它们应该是语义相似的,这个相似度由连接它们的边的权值决定.网络表示可以保留节点之间的这种相似性信息,使得具有局部特性的节点在低维空间相互靠近.

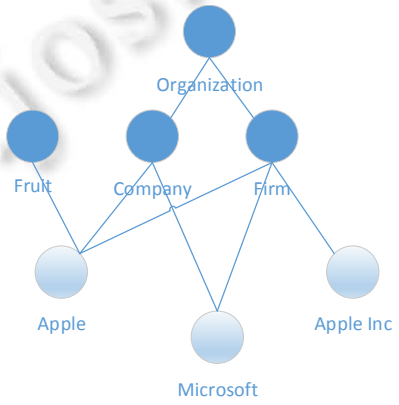


Fig.1 Samples of Probase network
图 1 Probase 网络的简单例子

下面介绍针对这种网络局部特性建模的网络表示算法[9].

对网络中每一条边(i,j),有一个联合概率公式表示节点 v_i 和 v_j 的一阶相似度:

$$p(v_i,v_j)=\frac{1}{1+\exp(-\vec{u}_i^T,\vec{u}_j)} \tag{3}$$

其中,$-\vec{u}_i^T\in R^d$ 是节点 v_i 的低维向量表示.而节点 v_i 和 v_j 的经验相似度定义为 $\hat{p}(i,j)=\frac{w_{ij}}{W}$,其中,$W$ 表示网络 G 中所有的边权总和.为了保留一阶相似度,优化以下目标函数:

$$O=d(\hat{p}(\cdot,\cdot),p(\cdot,\cdot)) \tag{4}$$

其中,$d(\cdot,\cdot)$表示两个分布的距离.使用 KL 距离公式替代 $d(\cdot,\cdot)$并忽略一些常数,可以得到:

$$O=\sum_{(i,j)\in E}w_{ij}\log p(v_i,v_j) \tag{5}$$

通过最小化目标函数 O,可以将网络中的每个节点表示到 d 维的向量空间中 $\{\vec{u}_i\in R^d\},i=1,2,...,|V|$.最后,可以得到一个词向量表示矩阵 U,包含知识图谱里每一个节点的词向量.

2.4 融合文本概念化与网络表示的观点检索模型

通过上述方法获得了文本概念化特征和网络表示特征之后,本文进一步使用上述表示特征进行观点检索方法的设计.考虑到在实际应用场景中,针对是否有人工标注的训练数据的不同情况,研究者往往使用不同的观点检索方式.其中,针对没有人工标注的训练数据的情况,研究者往往采用基于向量空间的统一观点检索方法;针对有人工标注的训练数据的情况,研究者往往采用基于排序学习模型的观点检索方法.因此,本节利用学习获得的特征表示设计了两种不同的观点检索方法,以应对实际应用场景中不同的情况与两种观点检索方法加以结合.

2.4.1 基于统一相关模型(unified relevance model)的观点检索方法

在以往的研究工作中,有以下统一检索模型,将文档的相关度得分视为文档观点得分的权值:

$$p(d\mid q,T)=\sum_i p(d\mid q,t_i)\propto\sum_i \alpha_i p(t_i\mid d,q)p(q\mid d)p(d)=I_{opn}(d,q,T)I_{rel}(d,q) \tag{6}$$

其中,d 表示一篇文档,T 表示观点词,q 表示用户查询文本,α_i 表示观点词的权重.公式(6)可以分为两部分:$\sum_i \alpha_i p(t_i\mid d,q)$ 表示观点得分,记为 $I_{opn}=(d,q,T)$;剩余部分表示文档与查询的主题相关度,记为 $I_{rel}=(d,q)$.

(1) 主题相关度计算

本文文档主题相关度采用 3 种方式计算:基于概念模型的文档主题相关度、基于网络表示的余弦相似度、概念模型和余弦相似度的线性加权.

文本经过概念化后得到 $d=(\langle c_1,w_1^d\rangle,...,\langle c_j,w_j^d\rangle,...,\langle c_k,w_k^d\rangle),j=1,2,...,k$ 和 $q=(\langle c_1,w_1^q\rangle,...,\langle c_j,w_j^q\rangle,...,\langle c_k,w_k^q\rangle)$,$j$=1,2,…,$k$.基于这种概念模型,本文用公式(7)表示推特与查询的主题相关度:

$$I_{rel}(d,q)=Sim_{BOC}(d,q)=\sigma\left(\sum_{c_i\in d\cap c_i\in q} w_i^d\cdot w_i^q\right),i=1,2,...,k \tag{7}$$

其中,$\sigma(x)$=1/(1+exp(−x))是一个 sigmoid 函数.

通过网络表示,本文将查询和文档都转换为低维空间中的向量,然后计算余弦值表示它们的相似度.查询的向量表示为 $\vec{q}=\dfrac{U^T\varphi(q)}{|\varphi(q)|}$,其中,$U$ 表示经网络表示得到的概念向量表示矩阵;$\varphi(q)$表示 q 涉及的概念集合,|$\varphi(q)$|表示 q 包含的概念个数.同样地,文档的向量表示为 $\vec{d}=\dfrac{U^T\varphi(d)}{|\varphi(d)|}$.最后计算这两个向量的余弦相似度:

$$I_{rel}(d,q)=Sim_{NE}(d,q)=\cos(\vec{d},\vec{q})=\frac{\vec{d}\cdot\vec{q}}{|\vec{d}\,||\,\vec{q}|} \tag{8}$$

此外,综合考虑概念模型和网络表示对实验性能的影响,本文还将公式(7)和公式(8)线性加权计算 $I_{rel}=(d,q)$:

$$I_{rel}=(d,q)=\lambda Sim_{BOC}(d,q)+(1-\lambda)Sim_{NE}(d,q),\lambda\in[0,1] \tag{9}$$

(2) 观点得分计算

观点得分 $I_{opn}=(d,q,T)$的计算除了沿用基于词典的观点挖掘方法[5]之外,本文还考虑了基于统计机器学习的观点挖掘方法来提高模型的观点泛化能力.使用 NBSVM[11]和 CNN[12]两种分类器对数据集进行主、客观分类,取主观性置信度作为观点得分.选用的训练语料是康奈尔(Cornell)大学提供的影评数据集,包含主、客观标签的句子各 5 000 句.

最后,用 $I_{rel}=(d,q)$为文档的观点得分 $I_{opn}=(d,q,T)$赋权,可得文档最终的观点检索评分公式为

$$Rank(d)=p(d\mid q,T)\overset{rank}{=}ScoreI_{opn}(d,q,T)\cdot ScoreI_{rel}(d,q) \tag{10}$$

2.4.2 基于排序学习模型 L2R 的观点检索方法

排序学习是一种数据驱动的方法,使用机器学习技术,根据带标签的数据和相关特征自动产生一个检索(排序)模型.本文使用 SVM_Rank[38]框架进行观点检索,这个框架在同类工作[35,39,40]中被广泛使用.SVM_Rank 将排序问题归结为二元分类问题.对于同一个查询 q,在其所有相关文档集里,任意两个不同标签值的文档,都可以得到一个训练实例 $((x_i^{(1)},x_i^{(2)}),y_i),i=1,2,...,m$,其中,$(x_i^{(1)},x_i^{(2)})$ 表示两个文档的特征向量,$y_i\in\{+1,-1\}$表示哪一个文

档应该排在前面,如果 $x_i^{(1)}$ 对应的标签值大于 $x_i^{(2)}$, 则 y_i 为+1,否则为–1.然后,优化以下损失函数:

$$\min_{\omega}\sum_{i=1}^{m}[1-y_i\langle\omega,x_i^{(1)}-x_i^{(2)}\rangle]_+ +\frac{1}{2C}\|\omega\|^2 \tag{11}$$

其中,$[x]_+$表示函数 max(x,0),C>0 是一个系数,m 表示文档对数,$\|\cdot\|$表示 L_2 范数.

本文选择的特征是基于排序学习的推特信息检索的常用特征[41]以及本文提出的文本概念化特征、网络表示特征和观点得分特征.

其中,基于排序学习的信息检索的常用特征可以从文档直接观察或者间接计算得到,这些特征在同类工作中均有被采用的实例,例如 BM25 得分、观点得分、是否含有链接、是否含有标签(#)和是否提及(@)他人、作者发布的推文数、关注作者的人数和作者关注的人数、作者被分组的次数等特征.

在此基础上,排序学习算法还采用了上文所提出的 3 种特征表示方法作为额外特征,包括文本概念化特征、网络表示特征、观点得分特征.其中,

- 文本概念化特征指的是文档包含的概念.本文利用概念生成算法从文档集生成概念词库.然后针对每篇文档,使用了类似词袋(bag-of-words)模型的表示方法,概念词库的每一个概念是否存在,均表示文档特征空间的一维;
- 网络表示特征指的是利用网络表示学习得到的网络表示作为特征,对于文档中涉及到知识库中的词语,利用网络表示结果将每个词语的表示累加求和并求取均值作为文档的特征.所以,这里的网络表示特征的维度与上文中利用网络表示学习获得的低维表示的维度相同;
- 观点得分特征是指利用前文提出的观点分析方法判断当前文档的观点得分.因此,此特征仅有一维.

使用上述特征,结合排序学习方法对待检索文档进行排序,并返回相应的检索结果.

3 实验结果与分析

3.1 数据集及评价指标

本文实验使用两个数据集:首先,根据 2014 年文献[6]的推特观点数据集进行实验,这一数据集共包含 49 个查询和 3 308 篇文档(在下文简称为推特 2014 数据集);由于这一数据集的数据量较少,本文利用推特提供的搜索结果及爬虫技术扩展数据集,共爬取 10 个查询的英文推特 29 634 篇.标注前,采用缓冲池(pooling)技术:针对每个查询,将本文检索模型和基准检索模型的各自检索结果中前 500 篇文档加入缓冲池,最后得到的缓冲池含有 7 172 文档.5 名标注人员对缓冲池中的文档进行二值标注,将与对应查询相关并且包含观点信息的文档标为 1,否则为 0.根据少数服从多数的原则对每篇文档进行判断,对缓冲池外的文档均标注为 0.下文将这一数据集记作扩展数据集.两个数据集的基本信息见表 2.

Table 2 Basic statistics of datasets

表 2 数据集基本信息

数据集	话题数量/个	文档数量/个	带观点的文档数量/个
推特 2014	49	3 308	590
扩展数据集	10	29 634	1 810

评价指标采用文本观点检索领域常用的 Mean Average Precision(MAP)、NDCG@10、R-precision(R-prec)和 binary Preference(bPref),具体计算公式如下:

$$MAP=\frac{\sum_{i=1}^{N^q}AP_i}{N^q},\ AP_i=\frac{1}{\sum_{i=1}^{N}r_i}\sum_{i=1}^{N}\frac{r_i\sum_{j=1}^{i}r_j}{i} \tag{12}$$

$$NDCG(n)=Z_n\sum_{j=1}^{n}(2^{r(j)}-1)/\log(1+j) \tag{13}$$

$$Rprec=\frac{\sum_{j=1}^{R}R_j}{R} \tag{14}$$

$$bpref = \frac{1}{R}\sum_{r} 1 - \frac{|n \text{ ranked higher than } r|}{R} \tag{15}$$

公式(12)中,N^q 为查询的数量,N 为总的文档数量.若第 i 个文档为带观点的主题相关文档,则 r_i=1;否则,r_i=0.

公式(13)中,Z_n 为标准化因子,用理想返回列表的 NDCG(n)作为因子进行归一化.$r(j)$指的是返回文档的评分,若相关,设为 2,否则,设为 1.

公式(14)中,R 为与查询相关并带有对查询观点的文档数量,R_j 为检索结果中第 j 个文档的评分,若是正确结果集中的文档,则取 1,否则,取 0.

公式(15)中,R 是与查询相关的文档个数,r 是具体的某一个相关文档,|n ranked higher than r|是排名比 r 靠前的非相关文档的数量.

上述评价指标中,MAP 是一个较为重要的指标,在本文的后续实验中,主要针对 MAP 进行不同方法之间的对比分析,其他 3 个指标仅作为参考指标.

3.2 实验对比

为了验证本文方法的有效性与普适性,本文分别与不需要训练语料的观点检索模型以及需要训练语料的排序学习模型进行对比.

- 观点检索模型

(1) SIGIR08[5]:基于词典的统一相关模型,通过二次组合方式将主题相似度得分与观点得分结合.该模型将观点词看成是均匀分布.模型使用传统信息检索方法 BM25 和基于词典的观点得分模型分别计算查询与文档的相关度和文档的观点得分;

(2) SIGIR08+Lexicon[6]:首先,基于异质图计算观点词在不同查询上的观点分布,然后将其融入 SIGIR08[5] 提出的模型中;

(3) SIGIR08_KG+Lexicon[10]:在文献[6]的基础上,使用知识图谱 freebase 的文本描述信息为用户查询进行查询扩展;

(4) BOC+X:本文方法,基于概念模型计算查询与文档的相关度并结合不同观点得分方法.有 3 种变形:BOC+Lexicon,BOC+NBSVM,BOC+CNN;

(5) NE+X:本文方法,基于网络表示计算查询与文档的相关度并结合不同观点得分方法.同样有 3 种变形:NE+Lexicon,NE+NBSVM,NE+CNN;

(6) BOC_NE+X:本文方法,基于概念模型和网络表示计算查询与文档的相关度,同情形(3)、情形(4),有 3 种变形:BOC_NE+BOC_Lexicon,BOC_NE+NBSVM,BOC_NE+CNN.

- 排序学习模型

(7) AAAI2012[39]:排序学习方法,利用推文特征、作者特征和观点特征训练排序模型;

(8) WWW2015[35]:排序学习方法,除了情形(7)提到的特征,加入了不同观点词典得到的观点特征、向量空间模型计算得到的查询相关特征和该推文发布的时长等特征;

(9) L2R+X:本文方法,除了信息检索的常用特征外,还加入了本文所提出的 3 种特征表示方法作为额外特征,包括文本概念化特征(BOC)、网络表示特征(NE)、观点得分特征(Lexicon,CNN 和 NBSVM),并在后续实验中分别对本文提出的特征进行组合分析.

基于上述方法,本文进行了以下 5 个实验.实验 1~实验 3 为统一相关模型的实验:实验 1 对比了本文方法与基准方法的实验结果.实验 2、实验 3 分析了本文统一相关模型方法中的参数设置.实验 4、实验 5 为排序学习模型的实验:实验 4 比对了本文特征与基准方法特征的实验结果,实验 5 分析了本文不同特征组合的实验结果.

实验 1:基于统一相关模型的观点检索对比实验.

为了验证本文提出的特征在统一相关模型的观点检索方法上的有效性,对比本文最优方法和基准方法在两个数据集上的实验结果.结果见表 3 和表 4.

Table 3 Comparison of our best approache and benchmark approaches on Tweet 2014 dataset (I)

表 3 本文最优方法与基准方法在推特 2014 数据集上的实验结果对比(I)

方法	MAP	NDCG@10	R-Prec	bPref
SIGIR08	0.330 6	0.406 8	0.405 1	0.410 8
SIGIR08+Lexicon	0.342 0	0.474 1	0.434 7	0.413 0
SIGIR08_KG+Lexicon	0.365 5	0.499 6	**0.447 7**	0.423 5
BOC_NE+NBSVM(本文方法)	**0.387 7**	0.502 6	0.444 9	**0.428 7**
BOC_NE+CNN(本文方法)	0.374 7	**0.523 6**	0.447 1	0.406 5

Table 4 Comparison of our best approache and benchmark approaches on extended dataset (I)

表 4 本文最优方法与基准方法在扩展数据集上的实验结果对比(I)

方法	MAP	NDCG@10	R-Prec	bPref
SIGIR08	0.317 7	0.425 8	0.278 5	0.277 1
SIGIR08+Lexicon	0.330 5	0.471 0	0.307 8	0.261 1
SIGIR08_KG+Lexicon	0.331 0	0.490 3	0.391 0	0.303 0
BOC_NE+NBSVM(本文方法)	**0.361 7**	0.504 1	**0.419 5**	**0.354 0**
BOC_NE+CNN(本文方法)	0.315 8	**0.524 2**	0.390 4	0.317 7

从实验结果可以看出,

(1) 首先比对两个数据集上的 3 种基准方法,SIGIR08 在 4 个指标上都是最低的,说明 SIGIR08+Lexicon 和 SIGIR08_KG+Lexicon 方法相比 SIGIR08 效果较优,因此,后续实验主要和这两种方法比对.在推特 2014 数据集(见表 3)和扩展数据集(见表 4)中,SIGIR08_KG+Lexicon 与 BOC_NE+NBSVM(本文方法)的实验结果均优于 SIGIR08+Lexicon,说明引入知识图谱分析用户查询和文档集的语义信息可以提高模型的知识泛化能力,进而提高原有观点检索的性能.并且需要注意的是:在不同的观点得分计算方式下,算法的性能存在一定的差异.主要原因在于:本文的观点得分是在其他领域的数据集上进行训练,对于观点检索所在领域的观点得分计算存在一定的误差.尤其是基于卷积神经网络的方法,由于需要大规模的训练样本,因此在小数据集上的性能受限,并不适合处理这一问题.最后,本文采用了基于 NBSVM 的观点得分计算方式;

(2) 在推特 2014 数据集中(见表 3),对比本文方法 BOC_NE+NBSVM 与 SIGIR08_KG+Lexicon,BOC_NE+NBSVM 优于 SIGIR08_KG+Lexicon,在 MAP、NDCG@10、bPref 指标上均有一定提升,分别提升了 6.1%、1.0%、1.2%.在扩展数据集中(见表 4),本文方法 BOC_NE+NBSVM 优于 SIGIR08_KG+Lexicon,在 MAP、NDCG@10、R-Prec、bPref 这 4 个指标上分别提升了 9.3%、2.8%、7.3%、16.6%.说明本文方法相比基于知识库的扩展,不仅可以有效分析用户查询的信息需求,同时可以准确理解文本集的信息,有效改善了传统的基于词袋模型的词匹配中词汇之间语义鸿沟的问题,提高了计算泛化能力,进而能够提高观点检索的性能.

实验 2:统一相关模型中不同特征组合的性能对比.

实验 2 比对不同观点得分计算方法结合不同相关度得分对检索性能的影响.实验结果见表 5 和表 6.

- 首先观察相同主题相关度得分结合不同观点得分的差异性.

在推特 2014 数据集中(见表 5),SIGIR08+NBSVM、BOC+NBSVM、NE+Lexicon、BOC_NE+NBSVM 分别取得相应的最高 MAP,在扩展数据集中(见表 6),SIGIR08+Lexicon、BOC+Lexicon、NE+Lexicon、BOC_NE+Lexicon 分别取得相应的最高 MAP,说明相同的相关度得分结合不同观点得分方法的检索效果具有明显的差异性.在基础数据集中,统计机器学习的观点得分优于基于词典的观点得分,但在扩展数据集中,统计机器学习的观点挖掘得分并未优于基于词典的观点得分.一方面是因为扩展数据集的测试语料远大于训练语料;另一方面是因为训练语料与测试语料的异质性,导致泛化能力不够.

- 然后观察相同观点得分结合不同主题相关度得分的差异性.

在两个数据集上,本文方法均取得相应的最高 MAP 值,说明本文方法能够有效提高模型的知识泛化能力和计算泛化能力,进而提高检索性能.同时还可以发现:在推特 2014 数据集中(见表 5),本文方法的 BOC_NE+

NBSVM 在 MAP 和 bPref 指标上均达到最好,BOC_NE+Lexicon 在 R-Prec 指标上获得最优值,BOC_NE+CNN 在 NDCG@10 指标上获得最优值.在扩展数据集中(见表 6),BOC_NE+Lexicon、BOC_NE+CNN、BOC_NE+Lexicon、BOC_NE+NBSVM 分别获得了 4 个指标的最优值,再一次验证了本文方法更加有效检索到与查询主题相关观点的文本,说明了本文方法能够有效提高检索模型的知识泛化能力和计算泛化能力,进而提高观点检索的性能.

Table 5 Comparison of different feature combinations in unified relevance model on Tweet 2014 dataset

表 5 统一相关模型中不同特征组合在推特 2014 上的实验结果对比

方法	MAP	NDCG@10	R-Prec	bPref
SIGIR08+Lexicon	0.342 0	0.474 1	0.434 7	0.413 0
SIGIR08+NBSVM	0.369 3	0.481 8	0.426 6	0.404 6
SIGIR08+CNN	0.362 7	0.494 6	0.441 2	0.394 9
BOC+Lexicon	0.380 8	0.501 2	0.424 9	0.425 6
BOC+NBSVM	0.382 2	0.501 2	0.433 5	0.425 6
BOC+CNN	0.361 1	0.493 0	0.442 2	0.397 4
NE+Lexicon	0.381 9	0.503 0	0.447 3	0.426 0
NE+NBSVM	0.378 3	0.503 1	0.437 0	0.425 9
NE+CNN	0.364 8	0.520 7	0.438 1	0.401 8
BOC_NE+Lexicon	0.384 8	0.503 5	**0.448 9**	0.428 5
BOC_NE+NBSVM	**0.387 7**	0.502 6	0.444 9	**0.428 7**
BOC_NE+CNN	0.374 7	**0.523 6**	0.447 1	0.406 5

Table 6 Comparison of different feature combinations in unified relevance model on extented dataset

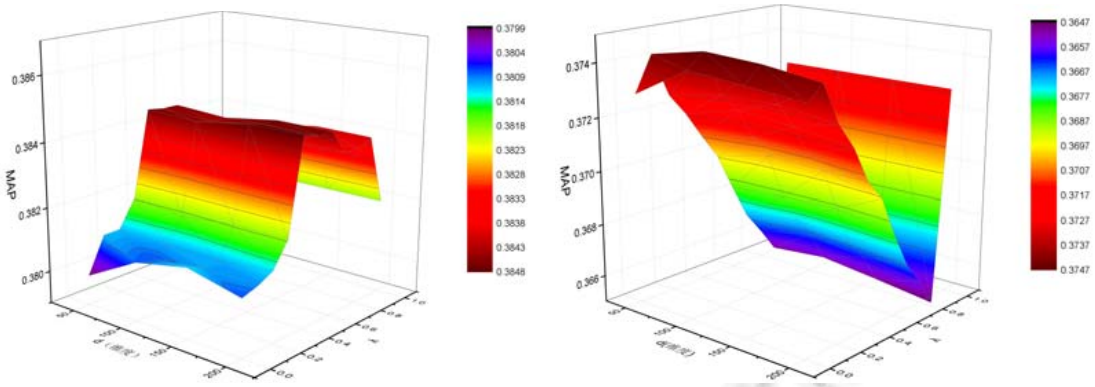
表 6 统一相关模型中不同特征组合在扩展数据集上的实验结果对比

方法	MAP	NDCG@10	R-Prec	bPref
SIGIR08+Lexicon	0.330 5	0.471 0	0.307 8	0.261 1
SIGIR08+NBSVM	0.281 2	0.481 4	0.287 2	0.238 7
SIGIR08+CNN	0.259 9	0.411 3	0.261 9	0.218 4
BOC+Lexicon	0.372 8	0.467 4	0.415 0	0.353 4
BOC+NBSVM	0.361 7	0.467 4	0.402 1	0.353 4
BOC+CNN	0.275 5	0.504 0	0.286 9	0.299 2
NE+Lexicon	0.374 1	0.490 8	0.410 6	**0.354 0**
NE+NBSVM	0.347 9	0.488 6	0.419 5	0.353 2
NE+CNN	0.304 6	0.504 9	0.384 1	0.314 3
BOC_NE+Lexicon	**0.374 7**	0.504 2	**0.421 6**	0.353 4
BOC_NE+NBSVM	0.361 7	0.504 1	0.419 5	**0.354 0**
BOC_NE+CNN	0.315 8	**0.524 2**	0.390 4	0.317 7

实验 3:统一相关模型中不同特征权重参数的对比实验.

本文模型 BOC_NE+Lexicon、BOC_NE+NBSVM、BOC_NE+CNN 均涉及两个参数:向量维度 d 和平滑参数 λ.由于在 4 个评估指标中 MAP 较为重要,因此本实验研究不同参数下对这 3 个模型 MAP 的影响.维度 d 设置为 50、100、150、200,平滑参数 λ 的范围为 0~1,步长为 0.1.λ 为 0 时表示基于网络表示的主题相关度得分,λ 为 1 时表示基于概念模型的主题相关度得分.实验结果如图 2~图 4 所示.

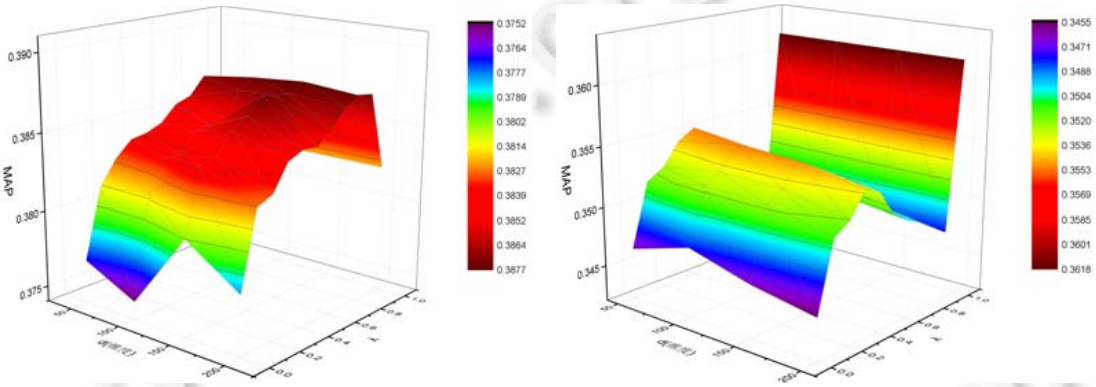
图 2(a)、图 3(a)和图 4(a)展示的是在推特 2014 数据集中,本文 3 种模型的 MAP 随参数 d 和 λ 的变化情况.BOC_NE+Lexicon(如图 2(a)所示)模型中,当 d 一定时,MAP 随着 λ 的增长而提升,在 λ 为 0.4 时达到峰值,随后又开始下降.其中,当 d 为 200、λ 为 0.4 时取得最大值.类似地,BOC_NE+NBSVM(如图 3(a)所示)模型当 d 为 150、λ 为 0.8 时 MAP 取得最大值,BOC_NE+CNN(如图 4(a)所示)模型在 d 为 150、λ 为 0.4 时 MAP 取得最大值.整体而言,在 λ 确定的前提下,网络表示的维度 d 对性能影响往往不大,体现为上述图表中维度方向的边际分布往往比较平缓.但是具体的 λ 则对性能的影响较大,即:文本概念化与网络表示的权重对于性能的影响较大.3 种不同的观点建模方法在两个数据集上的结果显示,这种参数设置对 MAP 的影响在 1%左右.考虑到表 3 与表 4 中本文提出方法相较于基准方法提升的幅度较大,所以这种参数设置并不影响本文方法相较于基准方法的性能优势.



(a) 推特 2014 数据集 (b) 扩展数据集

Fig.2 MAP of BOC_NE+Lexicon with different parameters on two datasets

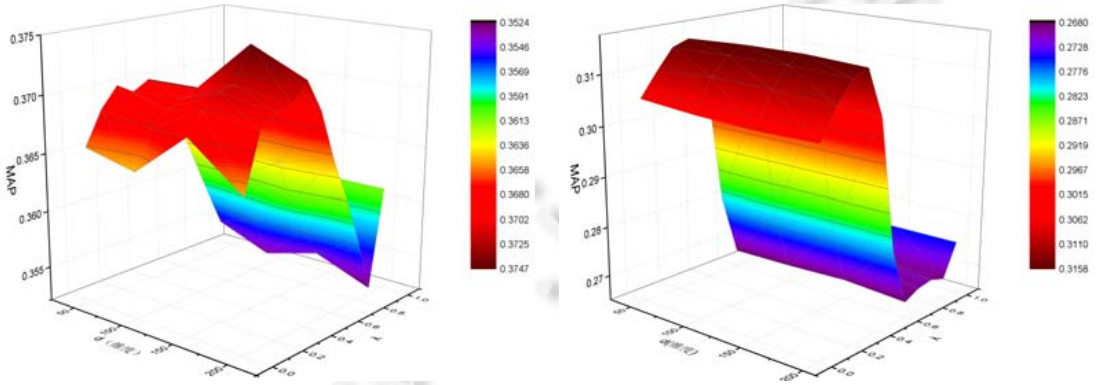
图 2 BOC_NE+Lexicon 在两个数据集中不同参数对 MAP 的影响



(a) 推特 2014 数据集 (b) 扩展数据集

Fig.3 MAP of BOC_NE+NBSVM with different parameters on two datasets

图 3 BOC_NE+NBSVM 在两个数据集中不同参数对 MAP 的影响



(a) 推特 2014 数据集 (b) 扩展数据集

Fig.4 MAP of BOC_NE+CNN with different parameters on two datasets

图 4 BOC_NE+CNN 在两个数据集中不同参数对 MAP 的影响

实验 4:基于排序学习模型的观点检索对比实验.

为了进一步验证文本概念化特征、网络表示特征和观点得分特征在观点检索任务中的性能,本文进一步构造了基于排序学习模型的观点检索实验,在信息检索的常用特征基础上加上本文提出的特征进行实验,选取其中最优的特征组合进行对比.实验结果见表 7 和表 8.

Table 7 Comparison of our best approache and benchmark approaches on Tweet 2014 dataset (II)

表 7 本文最优方法和基准方法在推特 2014 上的实验结果对比(II)

方法	MAP	NDCG@10	R-Prec	bPref
AAAI2012	0.432 0	0.587 9	0.491 2	0.607 6
WWW2015	0.434 5	0.592 8	0.502 9	0.601 3
L2R+BOC+NE+NBSVM(本文方法)	**0.442 0**	**0.607 9**	0.499 1	0.641 2
L2R+BOC+NE+CNN(本文方法)	0.441 7	0.599 1	**0.503 4**	**0.647 9**

Table 8 Comparison of our best approache and benchmark approaches on extended dataset (II)

表 8 本文最优方法和基准方法在扩展数据集上的实验结果对比(II)

方法	MAP	NDCG@10	R-Prec	bPref
AAAI2012	0.381 2	0.555 3	0.360 3	0.566 2
WWW2015	0.385 4	0.546 4	0.375 7	0.578 4
L2R+BOC+NE+NBSVM(本文方法)	**0.452 9**	**0.624 3**	**0.453 0**	**0.616 4**
L2R+BOC+NE+CNN(本文方法)	0.441 7	0.618 3	0.439 0	0.609 8

表 7 和表 8 显示:在推特 2014 数据集和扩展数据集中加入本文提出的 3 类特征后,MAP、NDCG@10、R-Prec 和 bPref 这 4 个指标均有提升.在推特 2014 数据集上(见表 7),4 个指标获得最优值的方法分别为 L2R+BOC+NE+NBSVM、L2R+BOC+NE+NBSVM、L2R+BOC+NE+CNN、L2R+BOC+NE+CNN.在扩展数据集中(见表 8),MAP、NDCG@10、R-Prec 和 bPref 这 4 个指标上达到最优值的方法均为 L2R+BOC+NE+NBSVM.说明 AAAI2012 和 WWW2015 提出的特征不够充分,产生的特征空间比较稀疏.而本文提出的利用知识图谱和网络表示产生的文档概念空间和文档低维向量能够缓解向量空间稀疏的问题.

实验 5:排序学习模型中不同特征组合的性能对比.

在基于排序学习方法的观点检索中,不同的特征组合导致的检索性能可能存在差异,实验 5 研究的是在信息检索的常用特征基础上加上本文提出的不同特征组合的观点检索性能.

表 9 和表 10 显示了本文提出的 3 类特征的不同组合在两个数据集上的实验结果.可以看出,单独加入 3 种观点得分特征对排序学习的性能影响不大.这是由观点分类的训练数据和测试数据的异质性导致的.同时可以看出:文本概念化特征和网络表示特征均能有效提升模型的性能,特别是在扩展数据集上,因为扩展数据集数据量大,出现歧义的实体现象较多,说明文本概念化特征和网络表示特征能够有效改善特征稀疏所引起的语义相似度计算困难现象.而在两个数据集上均显示出,同时使用文本概念化特征与网络表示特征可以进一步提升系统性能.这也说明两类方法在具体的使用过程中有一定的互补性.最后,在两个数据集上的所有特征组合中,MAP 值最高的均是 L2R+BOC+NE+NBSVM.这进一步证明了本文提出的方法能够有效解决上文提出的现有观点模型的 3 个局限性问题,从而提高检索模型的性能.

Table 9 Comparison of different feature combinations in learning to rank model on Tweet 2014 dataset

表 9 排序学习模型中不同特征组合在推特 2014 上的实验结果对比

方法	MAP	NDCG@10	R-Prec	bPref
L2R+BOC	0.438 9	0.602 1	0.508 0	0.647 9
L2R+NE	0.441 7	0.592 7	0.504 0	0.611 4
L2R+CNN	0.435 4	0.597 8	0.501 1	0.616 4
L2R+Lexicon	0.434 9	0.590 0	0.494 0	0.607 0
L2R+NBSVM	0.435 6	0.586 2	0.497 1	0.606 7
L2R+NE+CNN	0.440 2	0.596 5	0.507 2	0.620 2
L2R+NE+Lexicon	0.437 8	0.593 2	**0.518 2**	0.630 8
L2R+NE+NBSVM	0.435 6	0.587 6	0.488 1	0.610 2
L2R+BOC+CNN	0.435 0	0.599 4	0.506 1	0.650 7

Table 9 Comparison of different feature combinations in learning to rank model on Tweet 2014 dataset (Continued)

表 9 排序学习模型中不同特征组合在推特 2014 上的实验结果对比(续)

方法	MAP	NDCG@10	R-Prec	bPref
L2R+BOC+Lexicon	0.438 2	0.604 8	0.506 8	0.647 2
L2R+BOC+NBSVM	0.436 3	0.597 7	0.478 0	0.640 4
L2R+BOC+NE+CNN	0.436 4	0.599 1	0.503 4	0.647 9
L2R+BOC+NE+Lexicon	0.435 1	0.599 0	0.509 2	**0.652 1**
L2R+BOC+NE+NBSVM	**0.442 0**	**0.608 0**	0.499 1	0.641 2

Table 10 Comparison of different feature combinations in learning to rank model on extended dataset

表 10 排序学习模型中不同特征组合在扩展数据集上的实验结果对比

方法	MAP	NDCG@10	R-Prec	bPref
L2R+BOC	0.403 3	0.547 7	0.402 0	0.585 4
L2R+NE	0.384 2	0.581 6	0.392 8	0.564 0
L2R+CNN	0.381 0	0.534 8	0.354 1	0.576 2
L2R+Lexicon	0.383 0	0.565 5	0.366 2	0.576 0
L2R+NBSVM	0.384 8	0.534 7	0.356 5	0.562 9
L2R+NE+CNN	0.402 2	0.544 4	0.397 0	0.581 6
L2R+NE+Lexicon	0.403 5	0.546 5	0.396 1	0.583 8
L2R+NE+NBSVM	0.403 4	0.540 8	0.403 5	0.582 8
L2R+BOC+CNN	0.424 8	**0.687 2**	0.398 8	0.577 2
L2R+BOC+Lexicon	0.423 0	0.672 6	0.397 8	0.578 1
L2R+BOC+NBSVM	0.431 8	0.682 6	0.402 3	0.581 1
L2R+BOC+NE+CNN	0.441 7	0.618 3	0.439 0	0.609 8
L2R+BOC+NE+Lexicon	0.441 1	0.640 9	0.433 6	0.609 4
L2R+BOC+NE+NBSVM	**0.452 9**	0.624 3	**0.453 0**	**0.616 4**

4 总 结

本文提出了一种融合文本概念化与网络表示的观点检索模型.与现有研究工作不同,本文充分利用了知识图谱的结构化信息对用户查询和文本集进行语义分析,利用网络表示学习捕获知识图谱中节点之间的语义信息,利用统计机器学习的方法挖掘文本的倾向性信息.然后构建文本概念化特征、网络表示特征、观点得分特征这 3 类特征应用于统一观点检索模型以及基于排序学习的观点检索模型.实验结果表明:与现有工作对比,本文方法在 MAP 等指标上有明显的提升.在下一步工作中,首先可以进一步标注数据集,扩大训练集的语料,并结合常识性知识图谱,采用端到端(end to end)的模型进行训练,以期提高观点泛化能力.

References:

[1] Ounis I, Macdonald C, Rijke MD, Mishne G, Soboroff I. Overview of the TREC 2006 Blog track. In: Proc. of the 14th Text Retrieval Conf. (Trec 2006). Gaithersburg, 2006. 86−95.

[2] Pang B, Lee L. Opinion mining and sentiment analysis. In: Proc. of the Foundations and Trends in Information Retrieval. 2008. 1−135.

[3] Zhang W, Yu C, Meng W. Opinion retrieval from Blogs. In: Proc. of the 6th ACM Conf. on Information and Knowledge Management. ACM Press, 2007. 831−840. [doi: 10.1145/1321440.1321555]

[4] Wang XL, Wei F, Liu XH, Zhou M, Zhang M. Topic sentiment analysis in Twitter: A graph-based hashtag sentiment classification approach. In: Proc. of the 20th ACM Int'l Conf. on Information and Knowledge Management. ACM Press, 2011. 1031−1040. [doi: 10.1145/2063576.2063726]

[5] Zhang M, Ye X. A generation model to unify topic relevance and lexicon-based sentiment for opinion retrieval. In: Proc. of the 31st Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. ACM Press, 2008. 411−418.

[6] Liao XW, Chen H, Wei JJ, Yu ZY, Chen GL. A weighted lexicon-based generative model for opinion retrieval. In: Proc. of the Int'l Conf. on Machine Learning and Cybernetics. 2015. 821−826.

[7] Huang X, Croft WB. A unified relevance model for opinion retrieval. In: Proc. of the 18th ACM Conf. on Information and Knowledge Management. ACM Press, 2009. 947−956. [doi: 10.1145/1645953.1646075]

[8] Perozzi B, Al-Rfou, Skiena S. Deepwalk: Online learning of social representations. In: Proc. of the 20th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. ACM Press, 2014. 701−710.

[9] Tang J, Qu M, Wang MZ, Zhang M, Yan J, Mei QZ. LINE: Large-Scale information network embedding. In: Proc. of the 24th Int'l Conf. on World Wide Web, Int'l World Wide Web Conf. on Steering Committee. Florence, 2015. 1067−1077.

[10] Ma FX, Liao XW, Yu ZY, Wu YB, Chen GL. A text opinion retrieval method based on knowledge graph. Journal of Shandong University (Natural Science), 2016,51(11):33−40 (in Chinese with English abstract).

[11] Wang S, Manning CD. Baselines and bigrams: Simple, good sentiment and topic classification. In: Proc. of the Meeting of the Association for Computational Linguistics: Short Papers. 2012. 90−94.

[12] Kim Y. Convolutional neural networks for sentence classification. arXiv Preprint arXiv: 14085882, 2014.

[13] Robertson SE, Walker S, Jones S, Hancock-Beaulieu MM, Gatford M. Okapi at TREC-3. NIST Special Publication, 1995. 109−125. http://www.doc88.com/p-9972384819356.html

[14] Salton G, Wong A, Yang CS. A vector space model for automatic indexing. Communications of the ACM, 1975,18(11):613−620.

[15] Li X, Wang SG, Li DY, Kang XP, Zhai YH. Knowledge acquisition in incomplete information system based on formal concept analysis. Computer Science, 2014,41(7):250−253 (in Chinese with English abstract).

[16] Zhuang Y, Li GL, Feng JH. A survey on entity alignment of knowledge base. Journal of Computer Research and Development, 2016,53(1):165−192 (in Chinese with English abstract).

[17] Dalton J, Dietz L, Allan J. Entity query feature expansion using knowledge base links. In: Proc. of the 37th Int'l ACM SIGIR Conf. on Research & Development in Information Retrieval. ACM Press, 2014. 365−374. [doi: 10.1145/2600428.2609628]

[18] Xiong CC, Allan J. Query expansion with freebase. In: Proc. of the 2015 Int'l Conf. on the Theory of Information Retrieval. ACM Press, 2015. 111−120. [doi: 10.1145/2808194.2809446]

[19] Wang F, Wang ZY, Li ZJ, Wen JR. Concept-Based short text classification and ranking. In: Proc. of the 23rd ACM Int'l Conf. on Information and Knowledge Management. ACM Press, 2014. 1069−1078. [doi: 10.1145/2661829.2662067]

[20] Wang ZY, Zhao KJ, Wang HX, Wen JR. Query understanding through knowledge-based conceptualization. In: Proc. of the Int'l Conf. on Artificial Intelligence. 2015. 3264−3270.

[21] Li YH, Li DY, Wang SG, Zhai YH. Incremental entropy-based clustering on categorical data streams with concept drift. Knowledge-Based Systems, 2014,59(2):33−47.

[22] Zheng WG, Cheng H, Zou L, Jeffrey XY, Zhao KF. Natural language question/answering: Let users talk with the knowledge graph. In: Proc. of the 2017 ACM Conf. on Information and Knowledge Management. Singapore, 2017. 217−226.

[23] Hao YC, Zhang YZ, Liu K, He SZ, Liu ZY. An end-to-end model for question answering over knowledge base with cross-attention combining global knowledge. In: Proc. of the 55th Annual Meeting of the Association for Computational Linguistics. 2017. 221−231. [doi: 10.18653/v1/P17-1021]

[24] Liu ZY, Sun MS, Lin YK, Xie RB. Knowledge representation learning: A review. Journal of Computer Research and Development, 2016,53(2):247−261 (in Chinese with English abstract).

[25] Bordes A, Usunier N, Garcia-Duran A, Weston J, Yakhnenko O. Translating embeddings for modeling multi-relational data. In: Proc. of the NIPS. Cambridge, 2013. 2787−2795.

[26] Wang Z, Zhang J, Feng J, Chen Z. Knowledge graph embedding by translating on hyperplanes. In: Proc. of the AAAI. Citeseer, 2014. 1112−1119.

[27] Lin YK, Liu ZY, Sun MS, Liu Y, Zhu X. Learning entity and relation embeddings for knowledge graph completion. In: Proc. of the AAAI. 2015. 2181−2187.

[28] Ji GL, He SZ, Xu LH, Liu K, Zhao J. Knowledge Graph embedding via dynamic mapping matrix. In: Proc. of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th Int'l Joint Conf. on Natural Language Processing. 2015. 687−696. [doi: 10.3115/v1/P15-1067]

[29] Shi J, Gao H, Qi GL, Zhou ZQ. Knowledge graph embedding with triple context. In: Proc. of the 2017 ACM on Conf. on Information and Knowledge Management. Singapore, 2017.

[30] He SZ, Liu K, Ji GL, Zhao J. Learning to represent knowledge graphs with gaussian embedding. In: Proc. of the ACM Int'l Conf. on Information and Knowledge Management. 2015. 623−632.

[31] Ahmed A, Shervashidze N, Narayanamurthy S, Josifovski A, Smola AJ. Distributed large-scale natural graph factorization. In: Proc. of the 22nd Int'l Conf. on World Wide Web. ACM Press, 2013. 37−48.

[32] Jacob Y, Denoyer L, Gallinari P. Learning latent representations of nodes for classifying in heterogeneous social networks. In: Proc. of the 7th ACM Int'l Conf. on Web Search and Data Mining. ACM Press, 2014. 373−382.

[33] Santos RL, He B, Macdonald C, Ounis I. Integrating proximity to subjective sentences for Blog opinion retrieval. In: Proc. of the European Conf. on Information Retrieval. Springer-Verlag, 2009. 325−336. [doi: 10.1007/978-3-642-00958-7_30]

[34] Eguchi K, Lavrenko V. Sentiment retrieval using generative models. In: Proc. of the Conf. on Empirical Methods in Natural Language Processing. 2006. 345−354. [doi: 10.3115/1610075.1610124]

[35] Luo Z, Osborne M, Wang T. An effective approach to tweets opinion retrieval. World Wide Web, 2015,18(3):545−566.

[36] Kim YS, Song YI, Rim HC. Opinion retrieval systems using Tweet-external factors. In: Proc. of the 26th Int'l Conf. on Computational Linguistics (COLING), Proc. of the Conf. on System Demonstrations. Osaka: ACL, 2016. 126−130.

[37] Wang ZY, Cheng JP, Wang HX, Wen JR. Short text understanding: A survey. Journal of Computer Research and Development, 2016,53(2):262−269 (in Chinese with English abstract).

[38] Joachims T. Optimizing search engines using clickthrough data. In: Proc. of the 8th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. ACM Press, 2002. 133−142.

[39] Luo Z, Osborne M, Wang T. Opinion retrieval in Twitter. In: Proc. of the AAAI 2012. 2012. 507−510.

[40] Gerani S, Carman MJ, Crestani F. Investigating learning approaches for Blog post opinion retrieval. In: Proc. of the European Conf. on Information Retrieval. Springer-Verlag, 2009. 313−324.

[41] Duan YJ, Jiang L, Qin T, Zhou M, Shum HY. An empirical study on learning to rank of Tweets. In: Proc. of the 23rd Int'l Conf. on Computational Linguistics. Association for Computational Linguistics, 2010. 295−303.

附中文参考文献:

[10] 马飞翔,廖祥文,於志勇,吴运兵,陈国龙.基于知识图谱的文本观点检索方法.山东大学学报(理学版),2016,51(11):33−40.

[15] 李想,王素格,李德玉,翟岩慧.形式概念分析在不完备信息系统中的知识获取.计算机科学,2014,41(7):250−253.

[16] 庄严,李国良,冯建华.知识库实体对齐技术综述.计算机研究与发展,2016,53(1):165−192.

[24] 刘知远,孙茂松,林衍凯,谢若冰.知识表示学习研究进展.计算机研究与发展,2016,53(2):247−261. [doi: 10.7544/issn1000-1239.2016.20160020]

[37] 王仲远,程健鹏,王海勋,文继荣.短文本理解研究.计算机研究与发展,2016,53(2):262−269. [doi: 10.7544/issn1000-1239.2016.20150742]

廖祥文(1980－),男,福建安溪人,博士,副教授,CCF 高级会员,主要研究领域为文本倾向性检索与挖掘.

程学旗(1971－),男,博士,教授,博士生导师,CCF 会士,主要研究领域为网络科学与社会计算,互联网搜索与挖掘.

刘德元(1992－),男,硕士生,主要研究领域为知识图谱,观点检索.

陈国龙(1965－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为计算智能,计算机网络.

桂林(1987－),男,博士,主要研究领域为自然语言处理.