文章编号:1000－2367(2019)06－0022－09 **DOI**:10.16366/j.cnki.1000－2367.2019.06.004

一种融合 Wikipedia 类图和主题特征的短文本检索方法

李璞[1],肖宝[2],孙玉胜[1],张志锋[1],邓璐娟[1]

(1.郑州轻工业大学 软件学院,郑州 450000;2.北部湾大学 电子与信息工程学院,广西 钦州 535000)

摘 要:社交网络的快速发展催生出大量短文本数据.鉴于短文本具有长度短、信息量少、特征稀疏、语法不规则等特点,根据 Wikipedia 类图(Wikipedia Category Graph,WCG)中包含的结构信息,通过分析其中的主题特征,提出一种语义特征选择及关联度计算方法.以此为基础,通过计算用户查询与目标短文本之间的语义关联度,实现对短文本的检索和排序.最后通过在 Twitter 子集上的实验结果表明,融合 Wikipedia 类图和主题特征的短文本检索方法比现有一些检索方法在评估指标 MAP,P@k 及 R-Prec 上具有更好的效果.

关键词:Wikipedia 类图;主题特征;短文本;信息检索

中图分类号:TP391 **文献标志码**:A

随着新型社交媒体(如短信、微博和微信等)的普及,短文本已成为人们发布信息、进行社交活动的主要工具,同时也是政府、企业及时公开发布信息的重要媒介.与传统的长文本不同,短文本的内容组织通常不遵循语法结构且拼写随意新颖(如缩写、俚语).这些特点使得短文本的特征信息稀疏、所能表达的语义有限,从而机器很难在有限的语境中获取足够的信息量来对短文本进行理解和分析统计.这也导致传统信息检索技术无法有效地对其进行处理.针对上述问题,本文以语义关联度为出发点,将 Wikipedia 作为外部语义知识源,来研究短文本检索技术.根据 Wikipedia 类图(WCG)中包含的层次结构,通过分析特征概念间的主题信息,提出一种语义特征选择及关联度计算方法.在此基础上,提出一种融合 Wikipedia 类图和主题特征的短文本检索方法,并通过实验测试验证了该方法的可行性和有效性.

1 相关研究现状

早期的短文本检索技术主要以传统长文本检索方法为基础,采用的技术都是基于“词袋”模型的关键词匹配策略,如 tf-idf,BM25 和概率模型[1]等.然而基于“词袋”模型的检索方法常常忽略了词项的多义、同义和词项变体使用的问题,并且忽视了对短文本隐含语义的扩展,无法很好地解决短文本缺乏语义特征信息的问题,导致检索的效果不够理想.为了解决这些问题,许多研究开始尝试借助外部知识源对短文本进行信息扩展,从而更好地对短文本包含深层语义信息进行理解.

当前的短文本理解方法主要分为 3 种语义模型:隐性语义模型、半显性语义模型和显性语义模型[2].隐性语义模型将短文本解释为一个隐性向量,向量中各维度包含的信息只能用于机器处理.具有代表性的有隐性语义分析(LSA)[3]、超模空间模拟语言模型[4]以及神经网络语言模型[5]和段向量模型[6].与隐性语义模型不同,半显性语义模型将向量中的各维度映射为 1 个主题.主题通常是一组词或概念的集合[7].具有代表性的有以 LSA 为基础的 probabilistic LSA(PLSA)[8]以及隐式狄利克雷模型(LDA)[9].不同于上述 2 种模型,在显性语义模型下,短文本向量的各维度被解析为一系列具有明确语义的概念,从而便于对该向量进行理解

收稿日期:2019-01-14;**修回日期**:2019-05-15.

基金项目:国家自然科学基金青年科学基金(61802352);国家自然科学基金(61772210;61872439);郑州轻工业大学博士科研基金资助(0215/13501050015);郑州轻工业大学校级青年骨干教师培养对象资助计划(2018XGGJS006);钦州市科学研究与技术开发计划项目(20189903);广西高校中青年教师基础能力提升项目(KY2019KY0463).

作者简介(通信作者):李璞(1983－),男,河南开封人,郑州轻工业大学讲师,博士,研究方向为语义计算,大数据语义分析,E-mail:superlipu@163.com.

并做进一步的优化.常见的 2 种显性语义模型为:显式语义分析方法(ESA)[10]和概念化方法[11].

基于上述不同的短文本语义理解模型,人们提出了一些不同的短文本检索方法.最简单的方法采用传统的关键词匹配策略.但这类方法无法解决歧义概念的问题[12].为了克服上述局限性,一些研究采用基于上下文中潜在语义以及时间分布信息和主题特征的策略对短文本进行检索[13-15].此外,与本文基于隐性主题语义模型的研究思路不同,在作者之前针对短文本的研究工作中,通过分析 Wikipedia 中的显性概念特征,提出一种显性语义模型下的短文本理解和检索方法[16].近两年,国内外对于短文本的研究和应用又取得了一些新的进展.刘德喜等[17]结合多重增强图和 LDA 模型,于 2018 年提出了一种面向社交短文本的检索方法.Chu 等[18]应用主题传播策略,提出了一种新的短文本聚类方法.Li 等[19]针对短文本预处理问题,提出一种新的噪声过滤方法,从而提高了短文本建模和检索的正确性.Chen 等[20]于 2019 年对基于 LDA 和 NMF 的短文本主题发现和检索策略进行了研究和实验分析,并指出 LDA 模型可以有效地对短文本进行主题建模,并提高短文本检索的效果.

然而现有方法普遍认为所有的概念都是独立和等价的,没有考虑概念之间存在的隶属关系[21].因此这种横向的语义扩展仍然无法很好反映短文本中包含的隐含信息.为此,本文通过引入 Wikipedia 类图,通过分析概念在 WCG 中的层次结构对短文本的语义特征进行纵向延伸,对短文本进行更加合理的语义模型构建,从而提高短文本的检索效果.

2　短文本语义关联度计算及检索模型

2.1　基于 Wikipedia 类图的短文本语义特征选择

针对文献[21]提出的问题,以 Wikipedia 为外部知识源,通过对现有 ESA 算法进行改进,获取前 k 个最相关的 Wikipedia 概念作为基本语义特征,从而对短文本进行显式语义模型构建.在此基础上,通过分析各特征概念在 WCG 中的层次结构,获取其父类和子类信息,对短文本的语义特征进行扩展,并以此作为后续基于主题特征的关联度计算方法的基础.

首先,给出几个重要概念的形式化定义及相关算法的关键步骤.

定义 1(相关概念列表,(Relevant Concepts List,$RL_{\text{Top-}k}$))　给定 1 个短文本 d,称 $L=\langle A_1,\cdots,A_k\rangle$ 为 d 对应的相关概念列表 $RL_{\text{Top-}k}$.其中 L 中的每一个元素 A_i 都是一个二元组,即 $A_i=\langle c_i,w_i\rangle$.其中 c_i 为数据源 Wikipedia 中的概念,w_i 为概念 c_i 对应的 tf-idf 权值,参数 $k=1,2,3\cdots$.对于列表中的任意 2 个元素 $A_i=\langle c_i,w_i\rangle$ 和 $A_j=\langle c_j,w_j\rangle$ 满足如下条件:① 若 $i\neq j$,则有 $c_i\neq c_j$;② 若 $i<j$,则有 $w_i\geq w_j$.

获取 $RL_{\text{Top-}k}$ 算法的关键步骤见算法 1.

算法 1:根据输入的短文本 d,从 Wikipedia 中返回 $RL_{\text{Top-}k}$

输入:给定 1 个短文本 d,1 个停用词列表 sl,1 个阈值 k;

输出:$RL_{\text{Top-}k}$.

具体步骤:

步骤 1　根据停用词列表 sl 消除 t 和 Wikipedia 概念所对应文章中的无用词汇;

步骤 2　使用词干提取算法对 d 和 Wikipedia 概念所对应文章中的有效词汇进行归一化;

步骤 3　对于 d,采用 tf-idf 方法为 Wikipedia 中的概念分配对应的权值;

步骤 4　构建 d 和 Wikipedia 中概念之间的倒排索引;

步骤 5　根据 tf-idf 权值对倒排索引中的概念进行排序;

步骤 6　$RL_{\text{Top-}k}$ 为步骤 5 中返回的有序倒排索引中前 k 个最相关的概念列表;

步骤 7　返回 $RL_{\text{Top-}k}$.

定义 2(WCG 类别集,(Category Set of WCG,CS_{WCG}))　给定 Wikipedia 概念 c,称 $C=\langle ca_1,\cdots,ca_n\rangle$ 为 c 对应的 WCG 类别集 CS_{WCG}.其中 C 中的每一个元素 ca_i 都是 WCG 中 c 的一个父类概念或子类概念.

下面通过一个例子对定义 2 进行直观的说明.

例 1　对于 Wikipedia 中的概念“Artificial Intelligence”,其在 Wikipedia 中相应的父类概念信息如图 1

所示,子类概念信息如图 2 所示.

图 1 “Artificial Intelligence”在 Wikipedia 中的页面及父类概念

Fig. 1 The page and supercategory concepts for “Artificial Intelligence” in Wikipedia

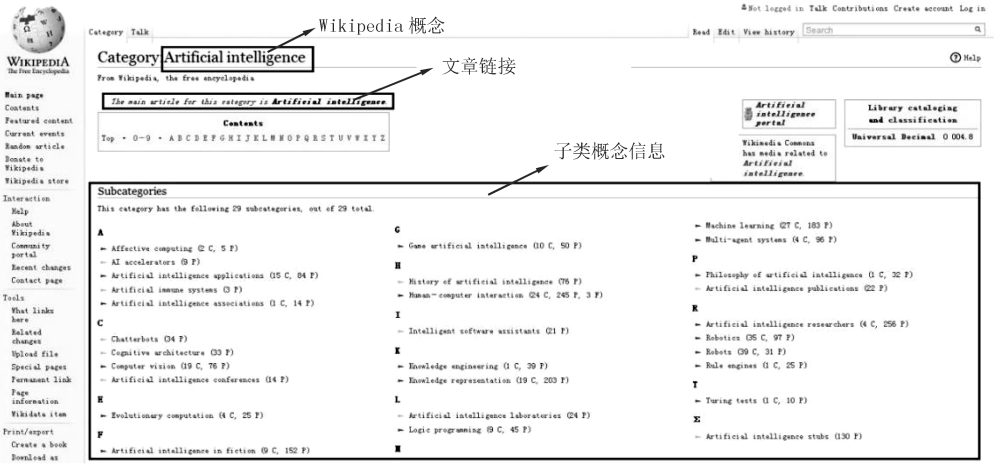

图 2 “Artificial Intelligence”在 Wikipedia 中的子类概念

Fig. 2 The subcategory concepts for “Artificial Intelligence” in Wikipedia

如果从图 1 和图 2 中抽取概念“Artificial Intelligence”在 Wikipedia 中的分类信息,可以得到对应的 CS_{WCG},其中包含了 7 个父类概念和 29 个子类概念.

根据定义 1 和定义 2,可以将 1 篇短文本转化为一个由最相关的前 k 个特征概念及其在 Wikipedia 中对应的类别集组成的特征向量空间,从而实现对短文本的语义特征选择.下面给出这个由相关概念的类别集组成的特征向量空间的形式化定义.

定义 3(相关类别集列表,(Relevant Category-Set List,$RCL_{\mathrm{Top}\text{-}k}$)):给定 1 个短文本 d,$L=\langle A_1,\cdots,A_k\rangle$ 为 d 对应的相关概念列表 $RL_{\mathrm{Top}\text{-}k}$,$C_i=\langle ca_1,\cdots,ca_n\rangle$ 为 L 中二元组 $A_i=\langle c_i,w_i\rangle$ 的特征概念 c_i 对应的 WCG 类别集 CS_{WCG}.称 $CL=\langle C_1,\cdots,C_k\rangle$ 为 d 对应的相关类别集列表 $RCL_{\mathrm{Top}\text{-}k}$,其中参数 k,$n=1,2,3,4,5\cdots$.

当给定 1 个目标短文本后,经过上述特征选择及类图信息抽取后,可以获得一个由相关概念的类别集组成的特征向量空间,这个向量空间中各维度间是按照特征概念与目标短文本之间的 tf-idf 权值(即 $RL_{\mathrm{Top}\text{-}k}$ 中二元组 $A_i=\langle c_i,w_i\rangle$ 中的 w_i)进行排序的.接下来,本文将在这种有序的特征向量空间中计算 2 个短文本间的语义关联度,从而实现短文本的检索.

2.2 基于主题特征的短文本语义关联度计算

众所周知,在向量空间模型下计算 2 个向量间的距离通常使用余弦度量公式.因此众多基于向量模型的信息处理方法也都采用余弦度量中的“点积”(dot product)公式来计算语义关联度(如:ESA,SVM 等).使用

余弦度量有一个非常重要的前提条件就是必须要保证 2 个向量具有相同的维度，同时 2 个向量的各个分量也要相同.

然而，由于在获取 $RL_{\text{Top-}k}$ 时引入了排序策略，因此对于 2 个不同的短文本〈d_1，d_2〉而言，其各自对应的 $RL_{(1)\text{Top-}k}$ 和 $RL_{(2)\text{Top-}k}$ 虽然具有相同的模（即特征概念数 k），但 2 个向量对应维度上的特征概念往往并不相同.不失一般性，在这种情况下使用余弦公式时，不得不将 2 个特征向量从原始大小扩展到二者的并集.这也正是传统基于向量空间模型的关联度算法必须构建高维度向量空间的原因.显然，这类算法存在 2 个问题：一是与短文本相关度很低（甚至无关）的特征概念参与计算，增加了算法的复杂度；二是因为 Wikipedia 的语料库非常大，通常情况下 1 个词只会在 Wikipedia 的很少一部分文章中出现.因此对于一个目标短文本而言，会生成一个高维稀疏的向量空间.显然使用余弦公式对这些 0 值分量的计算既占用大量时空资源，又没有任何意义.

针对上述问题，本节将研究如何在不对 $RL_{\text{Top-}k}$ 进行任何维度扩展的情况下，在相对低维的向量空间中实现短文本的语义关联度计算.

首先对于生成的 $RL_{(1)\text{Top-}k}$ 和 $RL_{(2)\text{Top-}k}$，需要对这 2 个向量空间中不同的分量进行分析，找出这些不同分量之间的对应关系.因此，有如下定义.

定义 4（$RL_{\text{Top-}k}$ 的关联系数）　给定 1 个短文本序对〈d_1，d_2〉，令 $L_1=\langle c'_1,\cdots,c'_k\rangle$ 和 $L_2=\langle c''_1,\cdots,c''_k\rangle$ 分别为 d_1 和 d_2 对应的 $RL_{(1)\text{Top-}k}$ 和 $RL_{(2)\text{Top-}k}$ 中的有序概念列表.则 $RL_{(1)\text{Top-}k}$ 和 $RL_{(2)\text{Top-}k}$ 之间的关联系数（association coefficient）可以被定义为 1 个 k 维向量，记为 $AC_{RL_{\text{Top-}k}}=\langle\lambda_1,\cdots,\lambda_k\rangle$，其中 $\lambda_i\in[0,1]$ 表示 $RL_{(1)\text{Top-}k}$ 和 $RL_{(2)\text{Top-}k}$ 对应分量上的概念 c'_i 和 c''_i 之间的距离或接近程度（$i\in\{1,\cdots,k\}$）.

接下来，我们将通过分析 $RL_{\text{Top-}k}$ 所对应的 $RCL_{\text{Top-}k}$ 中类别信息的主题特征，对 $AC_{RL_{\text{Top-}k}}$ 中的 $\lambda_i\in[0,1]$ 进行计算.

正如第 1 节研究现状中所阐述的，当前普遍采用的主题特征模型是隐式狄利克雷模型（LDA）.因此，本节应用 LDA 算法来计算 $RCL_{\text{Top-}k}$ 中类别信息的主题特征间的语义关联度，从而获取 $RL_{(1)\text{Top-}k}$ 和 $RL_{(2)\text{Top-}k}$ 的关联系数.

LDA 是一个基于概率的主题生成模型，可以通过无监督学习方式得到数据集所包含词及其主题的多项分布，其平滑版本的模型如图 3 所示.

图 3 中 M 是训练数据集中的文章总数，N 是文章的词的总个数，φ 是主题上的词分布，词 K 为主题总数，θ 是文章的主题分布，z 是每次生成文档词 w 时被选择的主题，因为 1 篇文档有多个主题，图 3 中的灰色框表示选择词 w 及其相关的主题 z 的步骤重复进行 N 次，α 和 β 是 2 个超参数，分别表示每篇文档的主题分布的先验 Dirichlet 分布以及每个主题的词分布的先验 Dirichlet 分布.对于 LDA 的详细介绍可以参见文献[9].

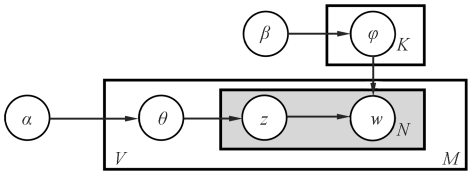

图 3　LDA 模型示意图

Fig.3　Graphical model representation of LDA

由定义 3 可知，$RCL_{\text{Top-}k}$ 中的每一个元素都是一个 WCG 类别集（见定义 2）.因此，在对 $RCL_{Top\text{-}k}$ 中的类别信息应用 LDA 进行主题特征建模时，仍然需要以 Wikipedia 为外部知识源对 LDA 进行训练.这里需要指出，根据信息熵的定义可知，在不同文本中出现次数过多或过少的词对文本的区分度较弱.因此，为了降低 LDA 的学习成本，在 LDA 的训练过程中，设定了 1 个值域把出现在不同文章中次数小于 20 及大于文章总数 10%的词项全部去除.然后通过 Gibbs 采样方法通过迭代计算得到 2 个概率矩阵 θ（document→topic）和 φ（topic→word）.θ 表示每个文档中每个主题出现的概率，φ 表示每个主题中每个单词的出现概率.

当 LDA 模型训练完成后，对于给定的〈d_1，d_2〉，根据定义 1 至 3 可将 d_1 和 d_2 对应的 $RL_{(1)\text{Top-}k}$ 和 $RL_{(2)\text{Top-}k}$ 转化为 $RCL_{(1)\text{Top-}k}$ 和 $RCL_{(2)\text{Top-}k}$，记为 $CL_{(1)}=\langle C'_1,\cdots,C'_k\rangle$ 和 $CL_{(2)}=\langle C''_1,\cdots,C''_k\rangle$.其中每个 C'_i 和 C''_i 都是 1 个 CS_{WCG}，即 $C'_i=\langle c'_{i1},\cdots,c'_{im}\rangle$，$C''_i=\langle c''_{i1},\cdots,c''_{in}\rangle$，其中 $m,n=1,2,3\cdots$.此时根据定义 4，每一个 $\lambda_i\in AC_{RL_{\text{Top-}k}}$ 可用如下公式定义：

$$\lambda_i=AssCoe(C'_i,C''_i),\tag{1}$$

其中 $\lambda_i \in [0,1]$ $(i \in \{1,\cdots,k\})$,函数 $AssCoe(C'_i,C''_i)$ 表示 $RL_{(1)\text{Top-}k}$ 和 $RL_{(2)\text{Top-}k}$ 对应位置上的特征概念在 WCG 中映射的 2 个 CS_{WCG} 之间的关联系数.

由图 1 和图 2 可知,在 WCG 中 C_i 的每一个 c_i 都对应着 1 篇 Wikipedia 文章 a_i.因此,$AssCoe(C'_i,C''_i)$ 可以使用如下 2 个公式进行表示:

$$AssCoe(C'_i,C''_i)=\max_{\forall(c'_p\times c''_q)\in C'_i\times C''_i}\{assCose(c'_p,c''_q)\},\tag{2}$$

其中 $1\leq p\leq|C'_i|$,$1\leq q\leq|C''_i|$,函数 $assCoe(c'_p,c''_q)$ 表示 2 个类别 c'_p 和 c''_q 之间的关联系数,并可表示为:

$$assCoe(c'_p,c''_q)=Sim(a'_p,a''_q),\tag{3}$$

其中 $a'_p(a''_q)$ 是类别(概念)$c'_p(c''_q)$ 在 Wikipedia 中对应的文章,函数 $Sim(a'_p,a''_q)$ 表示 2 篇文章之间的相似度.

此时,可以通过训练好的 LDA 模型将 a'_p 和 a''_q 映射到二维主题向量空间中.由于 LDA 生成的主题向量空间保存的是概率数据,因此需要使用针对概率数据的度量方法来计算 2 个概率分布之间的差异.当前度量概率分布的常用方法有 KL 散度(Kullback-Leibler divergence)、JS 散度(Jensen-Shannon divergence)、巴氏距离(Bhattacharyya distance)等,其中较为常用的是 KL 散度和 JS 散度.其中,JS 散度是 KL 散度的一种变形,解决了 KL 散度非对称性及无界性的缺点.因此,本文应用 JS 散度来计算(3)式中的 $Sim(a'_p,a''_q)$.JS 散度的计算方法如(4)式所示:

$$JSD(P\parallel Q)=\frac{1}{2}KLD(P\parallel\frac{P+Q}{2})+\frac{1}{2}KLD(Q\parallel\frac{P+Q}{2}),\tag{4}$$

其中 $KLD(P||Q)$为 KL 散度,计算方法如(5)式所示:

$$KLD(P\parallel Q)=\sum_{i=1}^{T}P_i\ln\frac{P_i}{Q_i}.\tag{5}$$

由(4)式可以看出,JS 散度的值域范围是[0,1],且 P 与 Q 相同时结果为 0,相反为 1.因此,需要对(4)式进行转换才能获得 $Sim(a'_p,a''_q)$合理的计算结果.具体公式如下

$$Sim(a'_p,a''_q)=1-JSD(P\parallel Q)=1-\frac{1}{2}(KLD(P\parallel\frac{P+Q}{2})+KLD(Q\parallel\frac{P+Q}{2})).\tag{6}$$

至此,可以通过(1)式至(6)式获得 λ_i 的计算结果,从而获得最终的 $AC_{RL_{\text{Top-}k}}$.对于给定的 2 个短文本 $\langle d_1,d_2\rangle$,通常情况下若 $d_1\neq d_1$,则有 $RL_{(1)\text{Top-}k}\neq RL_{(2)\text{Top-}k}$.对于 2 个不同的向量,在不求并集的前提下,仍然可以利用 $AC_{RL_{\text{Top-}k}}$ 将 $RL_{(2)\text{Top-}k}$ 中的特征概念转化为如下形式:$RL_{(2)\text{Top-}k}=\langle c''_1,\cdots,c''_m\rangle=\langle\lambda_1\cdot c'_1,\cdots,\lambda_m\cdot c'_m\rangle$,其中 $c''_i=\lambda_i\cdot c'_i$ 表明在 $RL_{(1)\text{Top-}k}$ 和 $RL_{(2)\text{Top-}k}$ 的相同分量位置上,对应特征概念 c'_i 和 c''_i 的关联程度.

以此为基础,对于 $\langle d_1,d_2\rangle$,可以在维度较低的 $RL_{(1)\text{Top-}k}$ 和 $RL_{(2)\text{Top-}k}$ 上定义一种新的语义关联度计算方法,具体公式如(7)式所示.

$$Rel(d_1,d_2)=\frac{\vec{V}(RL_{(1)\text{Top-}k})\cdot\vec{V}(RL_{(2)\text{Top-}k})}{|\vec{V}(RL_{(1)\text{Top-}k})|\cdot|\vec{V}(RL_{(2)\text{Top-}k})|}=\frac{\sum_{c'_i\in RL_{(1)\text{Top-}k}\cap c''_i\in RL_{(2)\text{Top-}k}}w_{(c'_i)}^{(d_1)}\cdot\lambda_i\cdot w_{(c''_i)}^{(d_2)}}{\sqrt{\sum_{c'_i\in RL_{(1)\text{Top-}k}}(w_{(c'_i)}^{(d_1)})^2}\cdot\sqrt{\sum_{c''_i\in RL_{(2)\text{Top-}k}}(w_{(c''_i)}^{(d_2)})^2}}.\tag{7}$$

2.3 短文本理解及检索模型构建

通过上面两节对短文本进行语义特征选择并通过分析特征概念在 WCG 中的主题信息,计算不同特征向量间的关联系数,可以在相对低维的语义空间下构建短文本的语义理解模型并应用(1)式至(7)式获得 2 个短文本间的语义关联度.接下来,将应用上述计算结果对短文本检索问题做进一步的研究.

从信息检索的角度看,用户输入的查询信息可以是 1 个关键词,也可以是短语或句子.但由于一般用户输入的查询内容不会很长,因此也可以将用户查询视为 1 条短文本信息.显然可以将用户输入的查询信息和目标短文本作为公式(7)中的 2 个变量,来计算查询信息和目标短文本之间的语义关联度,并根据语义关联度的大小对检索结果进行排序并将排序后的短文本列表返回给用户.本文提出的短文本检索模型如图 4 所示.

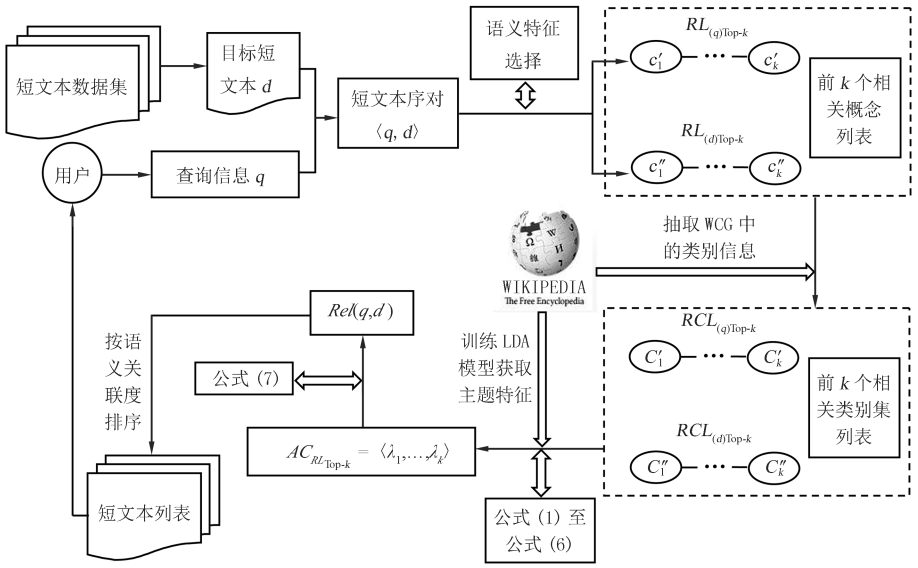

图 4　融合 Wikipedia 类图和主题特征的短文本检索模型

Fig. 4　Short text retrieval model combining Wikipedia category graph and topic features

3　实验与结果分析

3.1　Wikipedia 数据源及标准测试集

实验中使用的 2016 年 4 月 7 日发布的英文版 Wikipedia 数据①，并采用 JWPL (Java Wikipedia Library)②对 Wikipedia 中的语义数据进行预处理，从而在 WCG 中抽取类别及对应的文章信息.在预处理阶段，首先需要进行一些数据清洗工作 (如：去除 Wikipedia 间中包含有 File，Help，Draft 等的文件).

实验所采用的标准测试集来自文献[21]中收集的 Twitter 子集.该子集包含 3 980 061 个用户的属性参数以及用户之间的社交网络数据，收集了每个用户至少 600 条共 5 亿条内容种类多样的英语博文.为了与相关研究进行对比分析，本文采用了文献[21]中设计的 50 个用户查询.其中查询分为 20 个短文本查询和 30 个长文本查询.

这里需要注意的是，为了保证特征筛选及关联度计算过程的正确性和有效性，本文实验中构建了 1 个停用词列表(stop-words list)，并采用词干提取(又称词项归一)算法对 Wikipedia 页面及标准测试集中的停用词进行过滤并对词项进行规范化表示.

3.2　评价标准

由于在检索过程中引入了排序策略，所以实验中采用当前信息检索中普遍采用的 Mean Average Precision(MAP)、Precision at rank k(P@k)和 R-Prec 作为评价标准来衡量本文提出的短文本检索方法的有效性.3 种评价标准的详细介绍可参见文献[21].对应度量公式如下.

(1)Mean Average Precision(MAP)指的是在所有查询的平均正确率的均值，用 P_{Ma} 表示：

$$P_{Ma} = \frac{1}{N} \sum_{j=1}^{N} \frac{1}{Q_j} \sum_{i=1}^{Q_j} P(T_i), \tag{8}$$

其中 N 是查询的总个数，Q_j 是查询 j 返回的相关文档数，$P(T_i)$是相关文档在返回文档所在位置上的正确率.MAP 能直观反映系统在全部相关文档上性能的单值指标，查询得到的相关文档越靠前，其值就越高.

(2)Precision at rank k(P@k)指的是在返回查询结果的最前 k 条的正确率(本文实验中选取 $k=30$)用 P_k 表示：

① https://dumps.wikimedia.org/enwiki/20160407/enwiki20160407-pages-articles.xml.bz2

② https://www.ukp.tu-darmstadt.de/software/jwpl

$$P_k = \frac{\text{前 } k \text{ 个结果中与查询相关文档的个数}}{k}. \tag{9}$$

(3)R-Prec 表示检索出 R 篇文档时的正确率(Precision),用 P_R 表示.R 是与查询相关的文档总数,系统返回与查询相关的 R 个文档中共有 r 个文档是相互相关的:

$$P_R = \frac{r}{R}. \tag{10}$$

3.3 结果分析

实验中采用的主机为一台 16 GB 的内存和 4 核的 3.1 GB 的台式电脑.特征概念选择和 WCG 中类别信息抽取,以及基于 Gibbs 采样的 LDA 主题特征生成算法均通过 Java 语言实现.参考作者之前的研究结论[22]及实验结果对比,表 1 给出了取得较好实验结果时的各项参数值.

表 1 取得较好实验结果时的各项参数值

Tab.1 The parameter values for the better experimental results

算法阶段	参数名	取值
特征概念选择和 WCG 中类别信息的抽取	$RL_{\text{Top-}k}$ 的长度 k	10 000
主题特征生成	Gibbs 采样迭代次数	1 000
	主题数 k	1 000
	α	0.05
	β	0.001

按照表 1 中的参数设定,应用(8)式至(10)式,将本文提出的短文本检索方法与当前已有其他短文本检索方法在 MAP,P@k 和 R-Prec,3 个评价标准上的检索效果进行了比较.

正如在第 1 节研究现状中提到的,当前针对短文本理解和检索的诸多研究中采用了各种不同的策略.鉴于不同研究中使用的实验数据集不同,所以很难直接将各研究中的实验结果进行直接的对比和评价.因此,在保证实验数据一致性的前提下,为了更加清晰地分析本文算法的特点,选取了近两年几种相关性较高的短文本检索方法进行对比实验.同时,将其他相关研究中普遍采用的 2 种经典模型——ESA 和 LDA 应用于该标准测试集,从而更好地说明本文方法的有效性.表 2 对几种方法在标准测试集上相关评价结果进行了总结,其中第 1 行至第 4 行对应方法的实验数据可分别参见文献[21]和文献[15].

表 2 几种检索方法在不同查询需求上的评价结果

Tab.2 Evaluation results of several retrieval methods on different retrieval requirements

目标检索方法	短文本查询评价			长文本查询评价			研究时间
	MAP	R-Prec	P@k(k=30)	MAP	R-Prec	P@k(k=30)	
ESA	0.498	0.49	0.577	0.563	0.552	0.676	2007,2016
LDA	0.537	0.513	0.669	0.635	0.607	0.718	2003,2017
文献[21]的方法	0.550	0.531	0.695	0.688	0.674	0.764	2016
文献[15]的方法	0.553	0.542	0.721	0.733	0.71	0.854	2017
本文方法	0.613	0.586	0.762	0.755	0.741	0.859	2019

对表 2 中的各列结果进行纵向比较可以看出,与其他几种检索方法相比,本文提出的短文本检索方法在 2 类不同的检索实验中,所获得的检索结果在各项评价标准上均有所提高.可见通过抽取 Wikipedia 中特征概念的类图结构,对特征概念在 WCG 中的类别信息的主题特征进行分析后,可以获得更为相关的语义特征,并构建更加合理的短文本模型.当对表中 2 种检索需求进行横向比较时可知,所列 5 种检索方法对于长文本的检索效果都好于短文本的检索效果.这是因为长文本长度较短文本更长,自然也包含了更为丰富的信息.这种自身携带的原始语义信息往往要比人为扩充的语义信息更为精确,并且上下文之间具有更好的语义联系.可见文本长度对检索结果有着十分重要的影响.

在此基础上,继续对几种方法在标准测试集上的全部 50 个查询结果进行了综合的统计和分析,具体结果如表 3 所示,其中 Baseline 系统的实验数据也来自文献[21].可见本文提出的融合 Wikipedia 类图和主题

特征的短文本检索方法在各评价标准上都获得了一定的提高.

此外，从表 1 至表 3 可以看出，针对给定的标准测试集，尽管在构建特征空间时仅使用了 Wikipedia 中很少一部分概念($k=10\ 000$)作为向量空间(约占整个向量空间规模的0.18%)，本文提出的短文本检索方法在 MAP，R-Prec 和 P@k 上都能返回较好的评估效果.不仅如此，在允许更高计算复杂度的前提下，随着所构建显式语义空间中维度 k 的不断增加，本文提出的语义关联度计算方法将在 3 个评价标准上获得更好的评价结果.

表 3　几种检索方法的综合评价结果

Tab.3　The summary evaluation results of several retrieval methods

目标检索方法	综合评价			研究年份
	MAP	R-Prec	P@k($k=30$)	
ESA	0.539	0.52	0.627	2007，2016
LDA	0.602	0.595	0.693	2003，2017
Baseline 系统	0.593	0.582	0.64	2016
文献[21]的方法	0.638	0.601	0.719	2017
文献[15]的方法	0.657	0.632	0.765	2017
本文方法	0.711	0.673	0.816	2019

4　结束语

本文分析了传统信息检索方法在短文本检索时面临的局限性.根据 Wikipedia 中的语义信息，从 WCG 中抽取相关特征概念的类别信息以实现对短文本的特征选择及语义向量空间构建.在此基础上，通过分析类别信息在 Wikipedia 中对应的主题特征，对特征向量中对应位置的不同分量之间的关联系数进行计算，从而将 2 个不同的特征向量转化到相同的语义空间中.通过在这种较低维度的语义空间下计算用户查询信息和目标短文本之间的语义关联度，对短文本数据集进行排序并返回结果，从而实现短文本检索任务.由实验结果可知，本文提出的短文本检索方法可以获得更好的检索效果.在接下来的研究工作中，将进一步优化本文方法，构建更加合理的短文本理解及检索模型.

参　考　文　献

[1] Chen J，Nairn R，Nelson L，et al.Short and tweet：experiments on recommending content from information streams[C]//Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.New York：ACM，2010：1185-1194.

[2] 王仲远，程健鹏，王海勋，et al.短文本理解研究[J].计算机研究与发展，2016，53(2)：262-269.

[3] Deerwester S，Dumais S T，Furnas G W，et al.Indexing by latent semantic analysis[J].Journal of the Association for Information Science and Technology，1990，41(6)：391-407.

[4] Lund K，Burgess C.Producing high-dimensional semantic spaces from lexical co-occurrence[J].Behavior Research Methods，1996，28(2)：203-208.

[5] Bengio Y，Schwenk H，Sen Cal J S，et al.Neural Probabilistic Language Models[J].Journal of Machine Learning Research，2003，3(6)：1137-1155.

[6] Le Q V，Mikolov T.Distributed representations of sentences and documents[J].Computer Science，2014，4：1188-1196.

[7] Zhang H，Zhong G.Improving short text classification by learning vector representations of both words and hidden topics[J].Knowledge-Based Systems，2016，102：76-86.

[8] Hofmann T.Probabilistic latent semantic indexing[C]//Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval.New York：ACM，2004：56-73.

[9] Blei D M，Ng A Y，Jordan M I.Latent dirichlet allocation[J].Journal of Machine Learning Research，2003，3：993-1022.

[10] Gabrilovich E，Markovitch S.Computing semantic relatedness using Wikipedia-based explicit semantic analysis[C]//Proceedings of the International Joint Conference on Artifical Intelligence.San Francisco：Morgan Kaufmann Publishers Inc，2007：1606-1611.

[11] Wang Z，Zhao K，Wang H，et al.Query understanding through knowledge based conceptualization[C]//Proceedings of the International Joint Conference on Artificial Intelligence.San Francisco：Morgan Kaufmann Publishers Inc，2015：3264-3270.

[12] Tang J，Wang X，Gao H，et al.Enriching short text representation in microblog for clustering[J].Frontiers of Computer Science，2012，6(1)：88-101.

[13] Vicient C，Moreno A.Unsupervised topic discovery in micro-blogging networks[J].Expert Systems with Applications，2015，42(17)：

6472-6485.
[14] 韩中元,杨沐昀,孔蕾蕾,等.基于词汇时间分布的微博查询扩展[J].计算机学报,2016,39(10):2031-2044.
[15] 肖宝,李璞,胡娇娇,等.基于潜在语义与图结构的微博语义检索[J].计算机工程,2017,43(6):182-188.
[16] 李璞,张志锋,杨百冰,等.融合 Wikipedia 分类结构及显式语义特征的短文本检索[J].河南农业大学学报,2019,53(2):257-265.
[17] 刘德喜,付淇,韦亚雄,等.基于多重增强图和主题分析的社交短文本检索方法[J].中文信息学报,2018,32(3):110-119.
[18] Chu T Z,Cheng L,Wong H S.Corpus-based topic diffusion for short text clustering[J].Neurocomputing,2018,275:2444-2458.
[19] Li X,Yue W,Zhang A,et al.Filtering out the noise in short text topic modeling[J].Information Sciences,2018,456:83-96.
[20] Chen Y,Zhang H,Liu R,et al.Experimental explorations on short text topic mining between LDA and NMF based Schemes[J].Knowledge-Based Systems,2019,163:1-13.
[21] Kalloubi F,Nfaoui E H.Microblog semantic context retrieval system based on linked open data and graph-based theory[J].Expert Systems with Applications,2016,53:138-148.
[22] Li P,XIao B,Ma W J,et al.A graph-based semantic relatedness assessment method combining wikipedia features[J].Engineering Applications Of Artificial Intelligence,2017,65:268-281.

A short text retrieval method combining Wikipedia category graph and topic features

Li Pu[1],Xiao Bao[2],Sun Yusheng[1],Zhang Zhifeng[1],Deng Lujuan[1]

(1.Software Engineering College,Zhengzhou University of Light Industry,Zhengzhou 450000,China;
2.School of Electronics and Information Engineering,Beibu Gulf University,Qinzhou 535000,China)

Abstract: The rapid development of social networks has resulted in a large number of short text data. Considering the short length, little information, sparse features and irregular grammar, a semantic feature selection and relatedness computation method are proposed in this paper, which is based on the analysis of the topic features of the structural information contained in the Wikipedia category graph (WCG). On this basis, according to computing the semantic relatedness between user queries and the target short text, a short text retrieval and sorting method is realized. Finally, the experimental results on twitter subsets show that the short text retrieval method combining Wikipedia category graph and topic features outperforms other current retrieval methods on MAP, P@k and R-Prec.

Keywords: Wikipedia category graph; topic features; short text; information retrieval

[责任编校 陈留院 赵晓华]