doi:10.3969/j.issn.1001-2400.2013.02.015

面向文本分类的中文文本语义表示方法

宋胜利，王少龙，陈 平
(西安电子科技大学 软件工程研究所，陕西 西安 710071)

摘要： 为了解决词频统计文本表示方法中词语间语义信息缺失的问题，在考虑文本中词语上下文语境和语义背景信息的基础上，提出了一种新的中文文本表示模型——文本语义图. 该方法利用维基百科作为知识背景计算文本中实意特征词语的语义关联，将具有较强语义关系的词语合并成词包作为图的节点，节点权值用词包所包含词语的数目及词频计算；不同词包中词语间的上下文关系作为图的有向边，有向边权值用其邻接节点的最大权值表示. 该模型在较大程度地保留文本中词语上下文信息的同时强化了词语间语义内涵. 通过中文文本分类实验，文本语义图分类方法相对于支持向量机分类效率提升了 7.8%，同时错误率减少了 1/3，且表现出更好的稳定性. 实验结果表明在文本分类应用中，文本语义图模型能够有效地表示文本内容.

关键词： 分类；知识表示；相似度；文本语义图
中图分类号：TP391 **文献标识码：**A **文章编号：**1001-2400(2013)02-0089-09

Chinese text semantic representation for text classification

SONG Shengli, WANG Shaolong, CHEN Ping
(Research Inst. of Software Engineering, Xidian Univ., Xi'an 710071, China)

Abstract: Text representation based on word frequency statistics is often unsatisfactory because it ignores the semantic relationships between words, and considers them as independent features. In this paper, a new Chinese text semantic representation model is proposed by considering contextual semantic and background information on the words in the text. The method captures the semantic relationships between words using Wikipedia as a knowledge base. Words with strong semantic relationships are combined into a word-package as indicated by a graph node, which is weighted with the sum of the number and frequency of the words it contains. The contextual relationship between words in different word-packages is stated by a directed edge, which is weighted with the maximum weight of its adjacent nodes. The model retains the contextual information on each word with a large extent. Meanwhile, the semantic meaning between words is strengthened. Experimental results of Chinese text classification show that the proposed model can express the content of a text accurately and improve the performance of text classification. Compared to Support Vector Machines, Text Semantic Graph-based Classification can improve the efficiency by 7.8%, reduce the error rate by 1/3, and show more stability.

Key Words: classification; knowledge representation; similarity; text semantic graph

互联网技术的快速变革使得人类社会进入了信息极大丰富和快速更新的时代. 特别是近年来各种社交网络的出现，每天有海量文本信息不断产生和传播. 为了有效管理和利用这些电子化数据，文本挖掘和信息检索成为备受关注的研究领域. 文本分类是在预先给定的类别标记集合下，根据文本内容判定它的类别，广泛应用于自然语言处理与理解、内容信息过滤和舆情管理等多个领域. 文本表示是将自然语言文本描述为便于计算机

收稿日期：2011-11-11 **网络出版时间：**2012-11-16
基金项目：国家自然科学基金资助项目(JJ0500092301)；中央高校基本科研业务费资助项目(K50510230003)
作者简介：宋胜利(1981－)，男，讲师，博士，E-mail：shlsong@xidian.edu.cn.
网络出版地址：http://www.cnki.net/kcms/detail/61.1076.TN.20121116.0924.201302.109_032.html

处理的形式,它是文本分类处理及其他文本挖掘任务的基础和关键步骤.文本表示需要满足两个基本条件:首先文本表示过程中应保证文本语义信息的一致性,其次要求文本表示模型应便于进行后续计算过程.

文本表示方法按照结构特征可以分为 5 类:(1)集合理论.文本表示为特征词的集合.(2)代数理论.文本表示为向量、元组或矩阵.(3)概率统计.利用马尔可夫模型等将文本处理看做概率推理,能够考虑有限的组合关系.(4)图论.利用有向图描述文本概念之间的语义关系.(5)混合模型.使用最广泛的文本表示方法有两种:基于词频统计的向量空间模型(Vector Space Model, VSM)和基于语义分析的隐含语义索引(Latent Semantic Indexing, LSI).向量空间模型利用词袋(Bag Of Words, BOW)作为文本表示单元,将文档中包含的特征词看做多维的特征空间,每篇文档分别对应于该特征空间向量的一个实例.隐含语义索引利用词语与概念之间的映射关系,通过奇异值分析将文本中的索引词映射到低维空间中进行分析.

从自然语言理解的角度分析,英语是形合语言,造句要求词的形态变化符合规则,注重句法平面;而汉语是意合语言,造句要求词的意义搭配符合情理,注重语义平面.向量空间模型作为英文文本的一种有效表示方法,在中文文本表示中有一定的缺陷和不足[1]:(1)缺少词根特征,文本通常表示为一个高维度稀疏向量;(2)不同词语包含的信息熵及其对于文档主题的贡献度没有作区分;(3)中文词语丰富的含义使得语义相同或者相近的文档中相同的词语并不多,文本表示中丢失了概念之间天然的语义联系.隐含语义索引利用本体库或者概念词典实现词语的语义映射,其应用于中文信息处理中也会受到限制:(1)中文缺乏实用的语义词典;(2)这类模型通常过于复杂,其通用性受到限制,不便于进行后续计算.基于图结构模型的文本表示方法近年来成为研究热点,Schenker 等[2]首次将图结构引入到文本表示中,将 Web 文本中的英文特征项作为节点,以节点间的邻接共现关系为边进行构图,并用 3 种位置名称定义边的类别.这种方法在中文语境下无法直接使用,而且该模型构图时只考虑了边的位置信息,没有考虑特征项出现的频率及边的权重对文本表示效果的影响[3].

结合词语的语境和语义背景信息,笔者提出了一种面向文本分类应用的中文文本表示方法——文本语义图(Text Semantic Graph, TSG).文本语义图模型以文本中包含的实意词作为图的节点,词语在句子中的位置关系作为图的有向边,基于维基百科知识库分析词语之间的语义关系结合词频信息作为节点权重,有向边的权重表示了语境关系在文本中的重要程度.在研究了文本语义图模型构建方法及相似度计算模型的基础上,通过中文文本分类实验验证了该模型表达文本语义信息的有效性.基于文本语义图的文本分类方法在中文文本分类中相对于支持向量机分类方法具有更好的性能表现.

1 文本语义表示方法

大规模的文本文档中往往隐含了有价值的信息.如何自动发现文档中的潜在信息成为文本挖掘的一个重要的研究内容.文本语义表示方法的研究工作有助于提升文本挖掘任务的效率.由于向量空间模型在中文文本语义表示中的局限性,很多研究者通过实验试图找到能够精确表示文本语义内涵的表示方法,包括使用短语来代替词语、使用概念映射向量或者使用字符串核等,图是文本语义模型中最常见的表示形式.

基于图的文本语义表示方法主要是从 2000 年后开始的, Manuel、Aurelio 和 Alexander[4]在信息检索中提出了一些文本片段的表示和概念图匹配方法. Bhoopesh 和 Pushpak[5]提出了利用特征向量构建 UNL 图表示文本并结合 SOM 进行了聚类分析的方法. Schenker 等[2]在 2003 提出了用于网页聚类和分类的图结构文本表示模型,但是这个模型仅考虑了特征词之间是否共现而并未考虑共现的频率. Svetlana[6]提出了基于 VerbNet 和 WordNet 构建文本概念图.虽然这些模型能够体现出文本的语义信息,但是由于其结构过于复杂,没有一种有效的方法来计算图表示结构之间的相似度.

近年来,文本语义表示方法作为一个研究热点开展了大量的研究工作,并被广泛应用于各种不同的文本挖掘任务中. Song 和 Park[7]根据词语对句子含义贡献度的不同,提出了一种包含统计分析器、概念本体图表示和概念提取器文档的表示方法. Lee 等[8]针对领域本体构建方法研究了基于剧情构建文本本体模型,剧情包含文本中的概念属性和相关操作. Stavrianou 和 Andritsos[9]总结了文本语义表示模型并给出了比较分析,对后续的研究工作有很大的促进作用. Jin 和 Srihari[10]提出了一种基于图的文本表示结构,节点表示一

个特征概念,链接关系表示了概念之间的联系,链接的权重基于概念之间在同一个段落或句子中的共现率,利用骰子系数或极大似然估计的方法计算. Chang 等[11]利用类别标签作为原子概念,从维基百科词典中获取文本片段的显式语义分析(Explicit Semantic Analysis, ESA)[12]表示,构建带权向量来表示文本以便于进行后续计算. Li 等[13]认为基于词语在文本中出现的顺序对于文本主题的重要意义,通过计算在文本中出现的词的统计频率,然后按照词在文本词集中所占的比例,筛选出高频词和高频词义表示文本内容. Shaban[14]利用语义图模型作为文本的表示模式,分析句子的谓语结构并将结构中各个元素赋值,所有经过解析的句子合并后形成一个树结构表示文本的内容. Gad 和 Kamel[15]利用 WordNet 作为本体模型计算词项之间的语义关系,在表示文档时,加入了新的语义权重,在词频权重中引入了词项之间语义相似度的值,在语义上相关的词项被赋予更高的语义权重以强化文档所表示的语义中心. 国内关于文本语义表示方法的研究相对较少,Liu 等[16]利用词语网络描述文本结构信息,将词语之间的关系分为共现网络、句法网络和语义网络分别进行处理,主要利用了文本词语之间的结构关系以提升文本表示的准确性. 吴江宁等[3]提出了一种考虑词间语义和语序信息的基于图结构的中文文本表示方法,将文本特征项表示为图结构中的节点,特征项间的关系表示成节点间的有向边,提高了文本分类系统的性能,但在实际使用过程中,由于特征词数量太大,图结构往往很复杂,文本相似度计算时间消耗较高. 笔者提出的文本语义图模型同样采用了图结构作为文本表示模型,在表示过程中,不仅利用了词语语序结构,而且加入了词语之间的语义关系,根据词语的语义信息量筛选文本的语义特征,从而能够更有效地表示文本内容.

2　文本语义图模型及构建方法

2.1　文本语义图模型定义

文档中所出现的核心词汇之间在语义上相互关联,单个词语的含义在很大程度上依赖于其语境中的其他词语,甚至其语境可能会增强该词语在文档中的语义内涵. 如果将一个句子作为一个语义单元,那么该句子中的核心词语之间很可能具有较强的相关性. 文本语义图模型在文本表示中引入了词语语义关联及其上下文语境之间的联系.

定义 1　文本语义图. 文本语义图是一个有向图 G,定义为一个四元组 $G=(V,E,\alpha,\beta)$,其中 V 是一组节点的集合;$E\subseteq V\times V$,是一组有向边的集合;$\alpha:V\to\Sigma_V$,表示节点权值函数;$\beta:E\to\Sigma_E$,表示有向边权值函数.

定义 2　文本语义图的子图. $G_1=(V_1,E_1,\alpha_1,\beta_1)$,$G_2=(V_2,E_2,\alpha_2,\beta_2)$,当满足 $V_1\subseteq V_2$,$E_1\subseteq E_2\cap(V_1\times V_1)$,$\alpha_1(x)\leqslant\alpha_2(x)$,$\forall x\in V_1$,$\beta_1(x,y)\leqslant\beta_2(x,y)$,$\forall(x,y)\in E_1$ 时,G_1 称为 G_2 的子图,记为 $G_1\subseteq G_2$.

定义 3　文本语义图的阶. $G=(V,E,\alpha,\beta)$,其中节点的数目 $|V|$ 称为 G 的阶,记为 $|G|$.

定义 4　节点 v_i 的邻接边. G 满足 $(v_i,v_j)\in E$ 或 $(v_j,v_i)\in V$ 时,(v_i,v_j) 或 (v_j,v_i) 称为节点 v_i 的邻接边,节点 v_i 邻接边集合记为 Γ_i.

定义 5　节点 v_i 的 μ 词包. 表示节点 v_i 所包含的一个词语或者多个语义相似度超过阈值 μ 的词语集合,记为 $\Lambda_\mu(v_i)$. 词语 $w_i\in\Lambda_\mu$,当且仅当 $\{w_i\}=\Lambda_\mu$,或者 $S_{w_i,w_j}\geqslant\mu$,$\exists w_j\in\Lambda_\mu$.

2.2　词语语义相似度计算

词语语义相似度计算通常有 4 种方式:(1)利用词语在不同文本中的共现度计算;(2)利用词语的上下文语境计算;(3)采用潜在语义分析(Latent Semantic Analysis, LSA)方法;(4)利用知识库(如 WordNet、《知网》等)计算. 由于前 3 种方法只能在有限词集范围内使用,所以知识库是现在比较常用的方法. 中文信息处理中通常都采用《知网》计算词语之间的语义相似度,以《知网》中的语义关系为依据对词语概念中的义原进行分类,通过计算不同类型义原的相似度得到概念的相似度,取得了比较好的计算效果. 但《知网》在实际应用中有 3 个方面的不足:(1)概念的语义内涵受到义原的限制,很多专业词汇无法用义原来解释;(2)概念集的更新速度慢,很多新词的计算需要借助于其他扩展词典(同义词词林等)来完成;(3)词语相似度需要利用词图的相似性计算实现,计算过程时间较长.

维基百科提供了基于 Wiki 技术的大规模合作编辑环境,是迄今为止最大的本体库资源,词条之间的类

别结构和链接关系能够准确地反映概念之间复杂的语义相关性,可以被标注语义关系信息并自动生成机器可读的结构化语义知识. 文献[12]提出了基于维基百科的显式语义分析方法,相对于其他语义分析方法取得了更好的性能,并给出了不同计算方法间的性能比较(如表 1 所示).

表 1 词语语义相似度计算算法准确性比较

算法名称	作者/时间	准确性(相对人工)
WordNet	Jarmasz/2003	0.33～0.35
Roget's Thesaurus	Jarmasz/2003	0.55
Latent Semantic Analysis (LSA)	Finkelstein, et al./2002	0.56
WikiRelate!	Strube and Ponzetto/2006	0.19～0.48
Explicit Semantic Analysis (ESA)-Wikipedia	Evgeniy Gabrilovich, et al./2007	0.75
Explicit Semantic Analysis (ESA)-ODP	Evgeniy Gabrilovich, et al./2007	0.65

为了准确描述词语之间的语义关联,笔者利用基于维基百科的显式语义分析方法计算词语之间的语义相似度. 显式语义分析将维基百科数据集中每篇文档对应于一个词条,利用文档中的词语解释词条的语义内涵,词语的权值通过 TF·IDF 计算,每个词条就表示为一个带权向量. 然后按照词语建立倒排索引,每个词语可以表示为词条集对应多维空间中的向量,词语之间的语义相关性就可以通过向量距离进行计算.

文档 $T=\{w_i\}$,表示输入文档;$\langle v_i\rangle$ 表示与 $\{w_i\}$ 相对应的 TF·IDF 向量;$\boldsymbol{K}_i=\langle k_j\rangle$,表示词语 w_i 的倒排索引向量,其中 k_j 为词语 w_i 相对于词条 $c_j(c_j\in\{c_1,c_2,\cdots,c_N\})$ 的倒排权值,N 为维基百科中所有词条的数目. 文档 T 是长度为 N 的语义解释向量 $\boldsymbol{V}$,其第 i 维度词条 c_j 对应的词条权重为 $\sum_{w_i\in T}\boldsymbol{V}_i\cdot k_j$;文档 T_i 和 T_j 之间的语义相似度可以用其对应向量 $\boldsymbol{V}_i$ 和 $\boldsymbol{V}_j$ 夹角的余弦值表示;词语 w_i 和 w_j 之间的语义相似度可以用其对应向量 $\boldsymbol{K}_i$ 和 $\boldsymbol{K}_j$ 夹角的余弦值表示,即 $S_{w_i,w_j}=\boldsymbol{K}_i\cdot\boldsymbol{K}_j/(|\boldsymbol{K}_i|\times|\boldsymbol{K}_j|)$.

2.3 文本语义图构建方法

利用文本语义图模型表示文本文档时,每个节点对应于一个 μ 词包,节点权值表示其对应 μ 词包中包含的词语数目及出现频率;每个有向边表示其邻接节点对应的词语在句子中的位置关系,有向边权值表示有向边在文档中出现的次数. 文本语义图模型构建过程主要由两个阶段组成:

(1) 以句子为单位,采用了一种全切分与统计结合的分词算法,构造出基于统计词典的有向无环图,利用动态规划算法得到最佳的分词路径. 在句子分词过程中,根据词性对弱义词进行过滤,选择出名词或动词词性的词语作为核心词语. 文档 D 可表示为 $S=\{s_1,s_2,\cdots,s_K\}$,$s_i=(w_1,w_2,\cdots,w_T)$ 表示文档中第 i 条句子 S_i 中词性过滤后的核心词语列表,K 和 T 分别表示文档中句子的数目和句子中核心词语的数目.

(2) 基于句子核心词语列表,分别构建各条语句对应的文本语义图模型. 在基于第 1 条语句所构建的文本语义图基础上,分别依次将后续语句文本语义图合并进去. 先合并节点,计算新增节点与原节点之间的语义关系,如果节点之间词语相同或者语义相似度满足阈值条件,则将两个节点词语合并,节点权值相加,否则保留该节点;再合并有向边,如果新增有向边的相邻节点均被合并且合并后的节点之间存在有向边,则合并该两条有向边,有向边权值相加;最后,新合并节点的权值如果大于该节点邻接边的权值,则更新邻接边的权值为该节点的权值以强化节点之间的语义联系. 当所有句子均合并到第 1 条语句文本语义图模型后输出该模型,完成整个文档文本语义图表示模型的构造过程. 算法 1 描述了文本语义图的构建过程.

算法 1 文本语义图构建算法.

输入:句子序列 $S=\{s_1,s_2,\cdots,s_K\}$;语义相似度阈值 μ.

(1) 从 S 中取出第 1 条语句 $s_1=(w,w_2,\cdots,w_T)$,构建 G:$v_i=\Lambda_\mu(v_i)=\{w_i\}$,$e_i=(v_i,v_{i+1})$,$\boldsymbol{V}=\{v_1,v_2,\cdots,v_T\}$,$E=\{e_1,e_2,\cdots,e_{T-1}\}$,$\alpha(v_i)=1$,$\forall v_i\in\boldsymbol{V}$,$\beta(v_i,v_j)=1$,$\forall(v_i,v_j)\in E$;在 S 中删除 s_1.

(2) 从 S 中取出一条语句 $s_c=(w_1,w_2,\cdots,w_M)$,构建 G_c:$v_i=\Lambda_\mu(v_i)=\{w_i\}$,$e_i=(v_i,v_{i+1})$,$\boldsymbol{V}=\{v_1,v_2,\cdots,v_M\}$,$E_c=\{e_1,e_2,\cdots,e_{M-1}\}$,$\alpha_c(v_i)=1$,$\forall v_i\in\boldsymbol{V}$,$\beta_c(v_i,v_j)=1$,$\forall(v_i,v_j)\in E$;在 S 中删除 s_c.

(3) 从 V_c 依次取出节点 $v_i=\{w_i\}$, $i=1, 2, \cdots, M$, 分析其与 V 中任意节点 v_j, $j=1, 2, \cdots, T$, 的语义关系.

如果 $w_i \in v_j$, $\alpha(v_j)$ ++;

否则,如果 $S_{w_i, w \in v_j} \geqslant \mu$, $v_j=\Lambda_\mu(v_j) \cup \{w_i\}$, $\alpha(v_j)$ ++;

否则, $V=(v_1, v_2, \cdots, v_T) \cup \{v_i\}$, $\alpha(v_j)=1$.

(4) 从 E_c 依次取出有向边 $e_i=(v_i, v_{i+1})$, $i=1, 2, \cdots, M-1$, 分析其与 E 中任意有向边的关联关系, E 中必然存在两个节点 v_j, v_k 满足 $\Lambda_\mu(v_i) \subseteq \Lambda_\mu(v_j)$, $\Lambda_\mu(v_{i+1}) \subseteq \Lambda_\mu(v_k)$,

如果 $(v_i, v_k) \in E$, $\beta(v_j, v_k)$ ++;

否则, $E=\{e_1, e_2, \cdots, e_{T-1}\} \cup \{(v_j, v_k)\}$, $\beta(v_j, v_k)=1$.

(5) 如果 $|S| \neq 0$, 转(2).

(6) 在 G 中,对于每个节点 v_i,设置其邻接边集 Γ_i 中包含的每一条有向边的权值, $\forall (v_j, v_k) \in E$, $\beta(v_j, v_k)=\alpha(v_j) \geqslant \alpha(v_k)$? $\alpha(v_j)$: $\alpha(v_k)$.

输出:文本语义图 $G=(V, E, \alpha, \beta)$.

在构建文本语义图模型过程中,通过合并语句文本语义图模型中相同或语义相近的词语,并按照节点数目强化了有向边的权值,能够突出文档的语义内涵,便于描述文本中的隐含语义信息和主题特征.

2.4　文本语义图构建实例

通过一个新闻文本片段实例来说明具体如何构建文本对应的文本语义图模型.

文本 1　“日本财务省数据显示,日本 1 月未经调整贸易盈余年比减少 59.9%至 2 008 亿日圆,远弱于经济学家预期的减少 2.5%.”

(1) 经过中文分词和词性过滤,以逗号或者句号作为语句结束标志,获得了每条语句的核心词语列表. $s_1=$(日本,财务,省,数据,显示), $s_2=$(日本,月,调整,贸易,盈余,减少,圆), $s_3=$(经济学家,预期,减少).

(2) 根据算法 1 中所描述的文本语义图构建过程,输入 $S=\{s_1, s_2, s_3\}$, $\mu=0.7$(通过多次实验得到的经验值),根据词语语义相似度阈值合并后产生的 μ 词包如表 2 所示.

表 2　文本语义图节点信息

节点 v_i 编号	μ 词包($\mu=0.7$)	α 权值	说明
1	{日本}	2	单个词语
2	{财务,贸易}	2	S(财务,贸易) = 0.887 8
3	{省}	1	单个词语
4	{数据}	1	单个词语
5	{显示}	1	单个词语
6	{月,圆}	2	S(月,圆) = 1.0
7	{调整}	1	单个词语
8	{盈余}	1	单个词语
9	{减少}	2	单个词语
10	{经济学家}	1	单个词语
11	{预期}	1	单个词语

所构建文本语义图中包含的有向边及权值信息为

$\beta(v_1, v_2)=\beta(v_1, v_6)=\beta(v_2, v_3)=\beta(v_2, v_8)=\beta(v_7, v_2)=\beta(v_8, v_9)=\beta(v_9, v_6)=\beta(v_{11}, v_9)=2$,

$\beta(v_3, v_4)=\beta(v_4, v_5)=\beta(v_6, v_7)=\beta(v_{10}, v_{11})=1$.

最后输出文本 1 对应的文本语义图表示,用 NetDraw 绘制的文本语义图如图 1 所示.

3　文本语义图相似度计算

文本相似度计算是文本挖掘和信息检索的关键步骤,在词袋模型中一般采用“向量空间模型+余弦相似度”的模式进行计算.一种好的文档相似度计算方法应该能够反映文档中潜在的语义,并揭示隐藏在不同特

征词后面的相同概念语义.但是,余弦相似度方法要求特征词之间的语义关系必须是独立的,互不相交,忽略了基于语义层面的概念之间的语义关系,因此在计算文档语义相似度方面存在欠缺.笔者提出的文本语义图模型所表示的文档相似度计算中不仅考虑了词条集合的覆盖程度,而且考虑了两篇文档之间语义上的关联度.

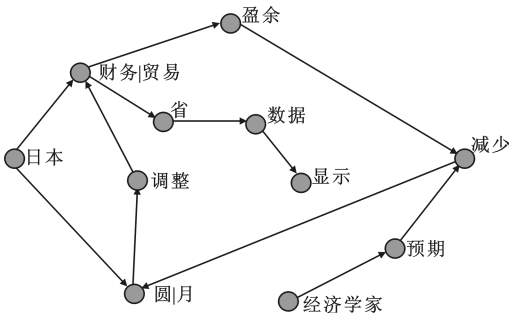

图 1 文本 1 构建的文本语义图结构

在文本挖掘类任务中,有两种类型的相似度计算需求——文本相似度和类别隶属度.前者计算两篇文本之间的相关程度,后者用于表示一篇文本相对于类别模型的相关程度.

$G_1=(V_1,E_1,\alpha_1,\beta_1)$ 和 $G_2=(V_2,E_2,\alpha_2,\beta_2)$,分别表示两篇文本文档;$G=(V,E,\alpha,\beta)$,表示经过机器学习后构建的类别文档模型,给出文本语义图相似度计算的相关定义.

定义 6 词包相似度.节点 v_i 和 v_j 的词包相似度用节点对应 μ 词包中所包含词语语义相似度的最大值表示,记为 $S^{\mu}(\Lambda_{\mu}(v_i),\Lambda_{\mu}(v_j))=M^{w_j\in\Lambda_{\mu}(v_j)}_{w_i\in\Lambda_{\mu}(v_i)}(S_{w_i,w_j})$.

定义 7 关联节点集.节点集 V_1 中与节点集 V_2 中任意节点 v_j 之间词包相似度大于等于 μ 的所有节点 v_i 所构成的集合称为 V_1 相对于 V_2 的关联节点集,它是节点集 V_1 的子集,记为 $U(V_1\rightarrow V_2)=\{v_i\mid S^{\mu}(\Lambda_{\mu}(v_i),\Lambda_{\mu}(v_j))\geqslant\mu;\ v_i\in V_1;\ v_j\in V_2\}$. V_1 相对于 V_2 的关联节点集中各个节点的权值之和可以表示为 $W(V_1\rightarrow V_2)=\sum_{v_i\in U(V_1\rightarrow V_2)}\alpha_1(v_i)$;$G_1$ 和 G_2 之间语义关联节点的权值之和 $W(V_1,V_2)=W(V_1\rightarrow V_2)+W(V_2\rightarrow V_1)$,$G_1$ 和 G_2 之间没有语义关联的节点集可表示为 $U^{-}(V_1\rightarrow V_2)=(V_1-U(V_1\rightarrow V_2))\cup(V_2-U(V_2\rightarrow V_1))$.

定义 8 关联边集.有向边集 E_1 中邻接点属于 V_1 相对于 V_2 的关联节点集的所有有向边所构成的集合称为 E_1 相对于 E_2 的关联边集,它是有向边集 E_1 的子集,记为 $R(E_1\rightarrow E_2)=\{(v_i,v_j)\mid(v_i,v_j)\in E_1;\ v_i,v_j\in U(V_1\rightarrow V_2)\}$. E_1 相对于 E_2 的关联边集中各条边的权重之和可以表示为 $W(E_1\rightarrow E_2)=\sum_{v_i\in R(V_1\rightarrow V_2)}\beta_1(v_i,v_j)$,两个文本语义图之间语义关联边的权值之和 $W(E_1,E_2)=W(E_1\rightarrow E_2)+W(E_2\rightarrow E_1)$,$G_1$ 和 G_2 之间没有关联的有向边集可表示为 $R^{-}(E_1\rightarrow E_2)=(E_1-R(E_1\rightarrow E_2))\cup(E_2-R(E_2\rightarrow E_1))$.

在上述定义的基础上,笔者给出了文本相似度和类别隶属度的计算公式.

(1) G_1 和 G_2 之间文本相似度用节点相似度和有向边相似度表示为

$$S(G_1,G_2)=\eta\frac{W(V_1,V_2)}{W(V_1,V_2)+|U^{-}(V_1,V_2)|}+(1-\eta)\frac{W(E_1,E_2)}{W(E_1,E_2)+|R^{-}(E_1,E_2)|}\quad,\tag{1}$$

其中节点相似度用 V_1 和 V_2 关联节点的权值之和在所有节点权值中所占的比例计算,有向边相似度用 V_1 和 V_2 关联边的权值之和在所有有向边中所占的比例计算,$\eta\in(0,1]$ 是权重调节因子.

(2) G_1 相对于类别 G 的类别隶属度用节点相似度和有向边相似度表示为

$$D(G_1,G)=\eta\frac{W(V_1\rightarrow V)}{\sum_{v_i\in V_1}\alpha_1(v_i)}+(1-\eta)\frac{W(E_1\rightarrow E)}{\sum_{(v_i,v_j)\in E_1}\beta_1(v_i,v_j)}\quad,\tag{2}$$

其中节点相似度用 V_1 相对于类别 V 的关联节点权值之和在 V_1 所有节点权值之和中所占的比例计算,有向边相似度用 E_1 相对于类别 E 的关联边权值之和在 E_1 所有有向边权值之和中所占的比例计算,$\eta\in(0,1]$ 是权重调节因子.

对于单类别文本分类中的类别隶属度,以定理形式给出 3 种特殊条件下的类别判断规则.

定理 1 如果 G_1 是且仅是 G 的子图,即 $D(G_1,G)=1$ 且 $\neg\exists G'$, s.t. $D(G_1,G')=1\wedge G'\neq G$,则 G_1 对应的文档分类到 G 对应的类别.

证明 文本分类是将待分类文本分类到其最大可能所属的类别中,即对于文档 d 和类别 C_i,$d\in C_i$ 当且仅当 d 对应的文本语义图表示 G_1 与类别 C_i 对应的文本语义图表示 G 的隶属度 $D(G_1,G)$ 为 d 和所有类别隶属度中的最大值.由于 $D(G_1,G)=1$,且 $\neg\exists G'$, s.t. $D(G_1,G')=1\wedge G'\neq G$,则类别 C_i 是文档 d 最大可能所属类别,即 $d\in C_i$.

定理 2　如果类别隶属度值相等且非 0,即 $D(G_1,G)=D(G_1,G')\neq 0$,则 G_1 对应的文档优先分类到 $|G|$ 与 $|G'|$ 值较小的类别.

证明　在文本分类过程中,文档与多个类别之间的隶属度相等且均为其最大值,即 $D(G_1,G)=D(G_1,G')\neq 0$,则需要计算这几个类别与文档之间的相似度并取其相似度大者.由相似度计算式(1)和式(2),当节点相似度和有向边相似度值相等时,文本语义图中所包含的节点数目(即图的度)越大,其隶属度的值越小.故根据文本分类的定义,G_1 对应的文档优先分类到 $|G|$ 与 $|G'|$ 值较小的类别.

定理 3　如果类别隶属度均为 0,即 $\neg\exists G'$, s.t. $D(G_1,G')\neq 0$,则 G_1 对应文档分类到 $M(|G|)$ 对应的类别.

证明　在文本分类过程中,文档与多个类别之间的隶属度均为 0 时,即 $\neg\exists G'$, s.t. $D(G_1,G')\neq 0$.由于硬分类要求必须将文档分到对应的类别,考虑到文本语义图中包含的节点数目越多,其与文档中存在相同词语的可能性越大,故 G_1 对应文档分类到 $M(|G|)$ 对应的类别.

在 2.4 节给出的实例基础上,计算文本 1 和文本 2 之间的语义图相似度以验证笔者提出的文本相似度计算方法.

文本 2　"日本财务省指出,日本对美国 1 月贸易盈余缩减 4.4%至 4 699 亿日圆.经济在 2004 年底确实走弱."

文本 2 对应的文本语义图表示结构如图 2 所示.

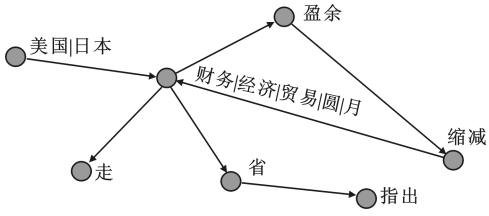

图 2　文本 2 构建的文本语义图结构

在 G_1 和 G_2 文本相似度计算过程中,设置参数 $\mu=0.7$,$\eta=0.5$,$W(V_1,\ V_2)=\ 10+\ 11=\ 21$,$W(E_1,\ E_2)=\ 12+\ 21=\ 33$,$|U^{-}(V_1,V_2)|=5+2=7$,$|R^{-}(E_1,E_2)|=6+2=8$.计算文本相似度 $S(G_1,G_2)=0.5\times(21/(21+7))+(1-0.5)\times(33/(33+8))=0.7775$,即说明两篇文本之间的相似度为 0.7775,符合人工理解和判断的结果.

4　文本分类实验及评价

中文文本分类实验数据集采用了复旦大学语料库 FDU 和谭松波分类语料 TSB.这些数据内容涵盖了经济、计算机、教育等 20 多个类别的近两万篇文本文档,在很多中文信息处理和文本挖掘中被用来作为标准评测数据集使用.

支持向量机(SVM)是一种基于统计学习理论的学习算法,该算法能较好地解决小样本学习[17]问题,广泛应用于当前主流的文本分类系统中并表现出较好的性能.笔者在分类实验中对比了两种类型的分类器——基于文本语义图的文本分类方法和基于向量的支持向量机分类方法.项目组在文本语义图分类方法基础上开发了文本语义图分类器,支持向量机分类实现了 SVM-light.实验中采用中科院中文分词系统 ICTCLAS 完成分词处理,特征选择选用信息增益,利用 TF-IDF 计算向量中的特征词权重.文本语义图分类器的语义相似度参数 $\mu=0.7$,文本语义图图相似度计算权重 $\eta=0.5$;SVM-light 使用默认参数设置,惩罚因子 $C=1\,000$,特征词数目设置为 1 000.详细实现过程描述如下:

(1) 根据表 3 中两类测试集所列类别,每个类别选择 10 篇带类别标签的文档作为训练样本进行分类模型训练.表 3 中同时列出了文本语义图分类器不同类别分类模型文本语义图的节点数目和支持向量机分类器不同类别实际使用的特征词数目.

(2) 从两个数据集所使用的类别中分别选择出 300～600 篇文档,分为 3 组,组成类别不平衡测试数据集,分别实施文本自动分类过程,具体每个类别所选出的文档数目在表 3 中给出.

(3) 利用支持向量机分类器对每个类别 3 组待分类数据集进行文本自动分类,将分类结果与待分类文档正确分类集进行比较,计算分类结果评价指标的平均值.

(4) 利用文本语义图分类器对每个类别 3 组待分类数据集进行文本自动分类,将分类结果与待分类文档正确分类集进行比较,计算分类结果评价指标的平均值.

文本分类实验所采用的评价指标体系包括准确率 P、召回率 R、宏平均 F 值 F_1 和错误率 E,计算方法如式(3)～(6)所示,其中,M 为测试样本集数目,T 表示属于该类别,F 表示不属于该类别,P 表示分到该类别,

N 表示未分到该类别.

$$P = TP/(TP + FP) \quad , \tag{3}$$

$$R = TP/(TP + TN) \quad , \tag{4}$$

$$F_1 = 2PR/(P + R) \quad , \tag{5}$$

$$E = (TN + FP)/M \quad . \tag{6}$$

表 3 测试数据集及分类模型中支持向量机特征词和文本语义图节点数目

FDU 测试集				TSB 测试集			
类别名称	文档数目	支持向量机特征	文本语义图节点	类别名称	文档数目	支持向量机特征	文本语义图节点
C11-Space	600	317	335	财经	300	453	423
C19-Computer	600	355	388	电脑	600	473	434
C3-Art	300	483	359	房产	300	384	337
C31-Enviornment	600	317	359	教育	300	413	384
C32-Agriculture	300	400	364	科技	300	407	396
C34-Economy	600	485	429	汽车	300	426	408
C38-Politics	600	437	396	人才	300	486	399
C39-Sports	600	418	418	体育	600	425	462
C6-Philosophy	300	451	349	卫生	300	369	358
C7-History	600	478	525	娱乐	300	447	560

基于上述评价指标体系,表 4 给出了支持向量机分类器和文本语义图分类器进行文本分类实验的性能表现.从实验结果可以看出,文本语义图分类器明显优于支持向量机分类器.宏平均 F_1 值在 6 次实验中最高提升了 13.07%(FDU-2),最低也提升了 4.33%(FDU-3);对应平均值分别在 FDU 和 TSB 数据集 3 次分类实验中提升了 8.41%和 7.25%,整体性能提升了 7.83%.错误率在 6 次实验中最高降低了 49.33%(FDU-2),最低降低了 26.40%(FDU-3);对应平均值分别在 FDU 和 TSB 数据集 3 次分类实验中降低了 37.47%和 31.29%,整体错误率降低了 34.38%.文本语义图分类方法相对于支持向量机分类方法保留了文本中词语的语境信息,并通过词包的引入和有向边权值进一步增强了词语之间的语义关联,更有利于分析和表示出文档所描述的主题信息.通过对这两类数据集进行的分类实验,验证了笔者提出的文本语义图在文本语义信息表达方面的有效性,而这些语义信息能够提升分类模型的准确性,并在文本相似度和类别隶属度计算中发挥重要作用.

表 4 文本分类实验性能指标对比分析

数据集	P		R		F_1		E		性能变化	
	支持向量机	文本语义图	支持向量机	文本语义图	支持向量机	文本语义图	支持向量机	文本语义图	F_1/%	E/%
FDU-1	0.8302	0.8888	0.8255	0.8840	0.8188	0.8829	0.0486	0.0299	+7.83	−38.48
FDU-2	0.7877	0.8864	0.7815	0.8770	0.7763	0.8778	0.0598	0.0303	+13.07	−49.33
FDU-3	0.8420	0.8780	0.8340	0.8670	0.8317	0.8677	0.0435	0.0328	+4.33	−26.40
TSB-1	0.8129	0.8783	0.8210	0.8700	0.8126	0.8706	0.0337	0.0236	+7.14	−29.97
TSB-2	0.8119	0.8801	0.8185	0.8655	0.8092	0.8691	0.0342	0.0235	+7.40	−31.29
TSB-3	0.8322	0.8924	0.8255	0.8715	0.8172	0.8762	0.0319	0.0215	+7.22	−32.60

分类方法的评估除了性能指标外,其稳定性表现也是影响分类器应用的一个重要因素.通过对分类结果进一步分析发现,TSB 语料相对于 FDU 语料在 3 次文本分类实验中表现稳定.在支持向量机分类中,分类的准确率为 81.9%±1.2%,召回率为 82.2%±0.4%,宏平均 F_1 值为 81.3%±0.4%,错误率为 3.3%±0.1%;在文本语义图分类中,分类的准确率为 88.4%±0.7%,召回率为 86.9%±0.3%;宏平均 F_1 值为 87.0%±0.3%,错误率为 2.3%±0.1%.文本语义图分类方法相对于支持向量机分类方法在各个指标值中相当或者更稳定,主要是由于文本语义图分类方法运用了图结构作为文本和分类模型的表示结构,在图节点

的构造过程中经过词性过滤,选取了句子中的名字和动词等实意词作为句子表示特征,这种特征相对于支持向量机特征向量中存在非实意词具有更好的稳定性.因此,笔者提出了文本语义图所采用的图结构以及构造过程中实意词的选择对于文本内容的描述具有积极的意义,加强了文本语义图分类方法的稳定性表现.

在表 4 中,FDU-2 和 FDU-3 的分类性能比较同样能够说明文本语义图分类方法相对于支持向量机分类的稳定性.以 FDU-2 作为基准进行比较,在支持向量机分类结果中,FDU-3 准确率从 0.787 7 到 0.842 0,变化率为 6.9%;召回率从 0.781 5 到 0.834 0,变化率为 6.7%;宏平均 F_1 值从 0.776 3 到 0.831 7,变化率为 7.1%;错误率的变化率为 27.3%.而在文本语义图分类结果中,FDU-3 相对于 FDU-2 的准确率、召回率、宏平均 F_1 值和错误率变化分别为 0.9%、1.1%、1.2%和 8.2%,文本语义图分类稳定性表现更为显著.

文献[3]采用复旦大学数据集的实验结果显示,基于图结构表示的文本分类方法相对于支持向量机分类其 F_1 值从 0.718 3 到 0.749 6,提高了 4.36%;笔者采用复旦大学数据集的分类,F_1 值提高了 8.41%,优于文献[3]的实验结果.文献[16]采用新浪新闻文本的实验结果显示,基于文本网络模型的文本分类方法相对于支持向量机分类其错误率从 29.2%到 11.34%,降低了 28.8%;笔者采用两种数据集的平均错误率降低了 34.38%,同样比文献[16]表现出更好的分类性能.实验结果比较说明:选取实意词作为语义图节点并计算词语之间的语义关系比直接利用统计特征词表示文本更有效.

5 总　　结

面对海量文本信息处理需求和当前中文文本信息处理中的语义表示问题,笔者提出了一种结合词语语境和语义背景信息的文本表示模型——文本语义图.利用基于维基百科的显式语义分析方法计算文本中词语之间的语义关系,将语义关系满足阈值条件的词语合并为词包,作为文本语义图的节点,将词语在文本中的语境关系作为文本语义图的有向边,通过设置节点和有向边的权值强化词语和文本之间以及词语与词语之间的语义联系.基于文本语义图表示模型,研究并提出了文本相似度和类别隶属度计算方法.最后通过两类中文文本分类数据集上的文本分类实验,验证了文本语义图在表示文本语义信息上的有效性.实验数据分析结果表明,基于文本语义图的文本分类方法相对于支持向量机分类具有更好的性能表现和稳定性,在整体性能平均提升了 7.8%的同时错误率减少了 1/3.

在目前工作的基础上,下一步工作是在文本聚类分析和语义信息检索等文本挖掘任务中引入文本语义图模型,作为文本表示结构,通过为文本处理提供语义信息以提升文本挖掘任务的执行效率.

参考文献:

[1] Li Yuhua, Mclean D, Bandar Z A, et al. Sentence Similarity Based on Semantic Nets and Corpus Statistics [J]. IEEE Trans on Knowledge and Data Engineering, 2006,18(8): 1138-1150.

[2] Schenker A, Last M, Bunke H, et al. Classification of Web Documents Using a Graph Model [C]//Proc of the 7th International Conference on Document Analysis and Recognition. Washington: IEEE Computer Society, 2003: 240-244.

[3] 吴江宁, 刘巧凤. 基于图结构的中文文本表示方法研究 [J]. 情报学报, 2010, 29(4): 618-624.
Wu Jiangning, Liu Qiaofeng. Research on Graph Structure Based Method for Chinese Text Representation [J]. Journal of The China Society for Scientific and Technical Information, 2010, 29(4): 618-624.

[4] Manuel M G, Aurelio L L, Alexander G. Information Retrieval with Conceptual Graph Matching[C]//Proc of the 11th International Conference on Database and Expert Systems Applications. London: Springer-Verlag, 2000: 312-321.

[5] Bhoopesh C, Pushpak B. Text Clustering Using Semantics [C]//Proc of the 11th International Conference on World Wide Web. New York: ACM Press, 2002: 79.

[6] Svetlana H. Construction of Conceptual Graph Representation of Texts [C]//Proc of Student Research Workshop at HLT-NAACL 2004. Stroudsburg: Association for Computational Linguistics, 2004: 49-54.

[7] Song W, Park S C. A Novel Document Clustering Model Based on Latent Semantic Analysis [C]//Proc of the 3rd International Conference on Semantics, Knowledge and Grid. Washington: IEEE Computer Society, 2007: 539-542.

(下转第 129 页)