文章编号：1673-5196(2015)04-0104-05

基于同义词词林扩展的短文本分类

王 东[1,2]，熊世桓[1,2]

(1. 贵州师范学院 数学与计算机科学学院，贵州 贵阳 550018；2. 贵州省高校工业物联网工程技术研究中心，贵州 贵阳 550018)

摘要：针对短文本特征稀疏导致的信息表示能力不足，提出基于同义词词林扩展的短文本分类方法. 该方法首先利用同义词词林确定短文本中主干词的同义关系，引入大规模词语搭配资源实现无指导多义词义项判别，从而确定候选扩展特征，最后计算候选扩展特征与给定上下文的语义关联性，将满足条件的候选特征扩展到特征向量中. 实验结果表明，该方法综合考虑的因素较全面，能够有效改善短文本的分类性能.

关键词：短文本分类；特征扩展；同义词词林；搭配词库

中图分类号：TP391 **文献标识码**：A

Short text classification based on synonymy expansion

WANG Dong[1,2], XIONG Shi-huan[1,2]

(1. Mathematics and Computer Science Institute, Guizhou Normal College, Guiyang 550018, China; 2. Industrial Internet of Things Engineering Research Center Education Institutions of Guizhou Province, Guiyang 550018, China)

Abstract: Aimed at the deficit of information expression ability caused by sparseness of short text feature, a method of short text classification is proposed based on synonymy expansion. In this method, a synonymy is employed to determine synonymous relation of main word in short text and large-scale word collocation resources are introduced to realize discrimination of unsupervised polysemous word so as to make the candidate expansion characteristics determined, Finally, by means of calculating the semantic relevance of candidate expansion feature to a given context, the candidate features meeting the conditions will be extended to the feature vector. The experimental result shows that in this method, more overall factors are taken comprehensively into account, so that a higher classification performance can be achieved.

Key words: short text classification; feature expansion; synonymy; collocation dictionary

新兴网络如即时通信、微博、论坛、电子邮件、网络视频、网络购物的广泛应用催生了大量的短文本数据，如搜索页面片断、聊天信息、微博信息、邮件主题、商品描述、观点评论、图片/视频文字介绍等. 这些文本长度很短，包含内容字数很少，且形式多样、数量庞大. 对如此庞大的海量文本数据，如何进行有效管理已成为广泛关注的问题.

文本分类技术作为一种有效的组织和管理数据的方法，在迅猛发展的网络中正发挥着举足轻重的作用. 然而，现有文本自动分类的研究与应用大多是针对长文本. 短文本因篇幅短、包含信息少、特征稀疏等特点使得用于长文本的分类技术难以对其进行简单移置. 解决短文本特征稀疏性问题的可行途径是对短文本进行信息补充，通过添加语义关联词扩展短文本，从而增强短文本表达的信息量[1-3]. 国内对此已进行了一些探索. 王细薇等提出一种基于特征扩展的中文短文本分类方法[4]，利用 FP-Growth 挖掘关联规则对短文本测试文档中的概念词语进行扩展. 宁亚辉等提出基于领域词语本体的短文本分类方法[5]，借助知网从语义方面将抽取的领域高频词扩展为概念和义元. 王盛等提出一种利用上下位关系的中文短文本分类框架[6]，利用“知网”确定训练文本中概念对的上下位关系，进而确定词语对的上下位关系，将其用于扩展测试文本的特征向量. 范云杰等提出基于维基百科的中文短文本分类方

收稿日期：2014-11-05

基金项目：贵州省优秀科技教育人才省长专项资金(黔省专合字(2012)82)，贵阳市科技计划项目(筑科合同[2013101]10-6号)

作者简介：王 东(1978-)，男，贵州贵阳人，副教授.

法[7],借助网络知识库维基百科抽取相关概念,对短文本的特征向量进行扩展. 袁满等提出一种基于频繁词集的特征扩展方法[8],挖掘具有相同类别倾向的频繁词集作为短文本特征扩展的背景知识库.

本文在该领域上做了进一步研究,利用词的同义关系引入同义词并将其扩展到短文本的特征向量中,在一定程度上弥补短文本信息量少的缺陷,实现分类性能的改善.

1　同义词词林的分析与处理

1.1　同义词词林

本文引入《同义词词林(扩展版)》[9]为语义资源. 该资源是一个具有 5 层的树状层次结构,前 3 层将词汇分成大、中、小三类,第 3 层根据词义的远近和相关性进一步细分为第 4、5 层,每层结构分别赋予了一个语义编码. 语义编码的第 1 层用大写英文字母表示,第 2 层用小写英文字母表示,第 3 层用二位十进制整数表示,第 4 层用大写英文字母表示,第 5 层用二位十进制整数表示,5 层结构构成的完整编码表示一个原子词群. 原子词群又通过 3 种标记"="、"#"、"@"分别表示同义词、相关词、只有一个词. 本文只选取标记为"="的原子词群.

《同义词词林(扩展版)》词典以文本文件形式存储,每一行表示一个原子词群. 文中对《同义词词林》的主要操作是获取给定词的同义词列表,直接利用原始词典效率很低,需要对词典进行重构,使之更加符合实际应用. 重构的同义词词典包括两个主要数据结构. 一是词向量,由所有原子词群中的词组成. 为提高查询速度,对词向量进行了排序,并建立索引;二是原子词群向量集,每条向量存储一个原子词群. 词向量通过指针关联对应的原子词群. 词典结构如图 1 所示.

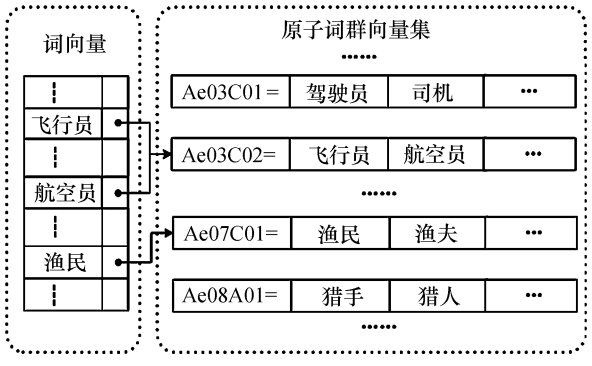

图 1　同义词词典

Fig. 1　Synonyms dictionary

要搜索指定词的同义词列表时,首先在词向量中搜索出指定词,再通过指向原子词群向量的指针获取同义词列表.

1.2　多义词义项判别

对《同义词词林》进行统计,发现词典中多义词是一个普遍的现象,词典中提取了表示同义词的原子词群中的总词数为 45 365 个,其中,多义词数有 10 479 个,占 23%. 当要进行扩展的特征项是一个多义词时,在词典中必然存在两个以上的原子词群,应该选择哪一个原子词群呢?这个问题实际上是一个词义标注问题,需要利用词义排歧技术根据句子的上下文判断出特征项的义项. 目前词义排歧的主流方法是基于语料库的统计方法. 根据训练语料事先是否经过人工标注又分为有指导和无指导两类[10]. 由于有指导方法需要大量人工标注工作,很难实现基于大规模标准语料的有指导词义排歧. 无指导的词义排歧方法不依赖人工标注语料,可实现跨领域大规模真实语料的训练和学习,能够有效克服数据稀疏问题[11]. 本文使用基于贝叶斯模型的无指导学习策略实现多义词义项判别. 首先将词义与上下文的关系形式化为一个二元组 $(s_i, V_{\text{context}})$. $V_{\text{context}}=(v_1, v_2, \cdots, v_n)$ 表示 s_i 的上下文词语的集合. 一般来说,上下文指一个句子. 贝叶斯模型将词义排歧看作一个词义分类问题,计算在特定上下文下多义词每个义项出现的概率,取具有最大条件概率的词义为判别结果. 公式表示为

$$\underset{s_i \in S}{\operatorname{argmax}}\, p(s_i) \prod_{v_k \in V}^{n} p(v_k \mid s_i) \tag{1}$$

其中:S 表示义项集合,v_k 为上下文词集合中的一个词. 公式(1)中,先验概率 $p(s_i)$和条件概率 $p(v_k \mid s_i)$ 并不能从语料中直接计算. 而是间接使用不同义项所属的原子词群在语料中的分布概率近似估算. 当语料规模足够大时,通过对大规模语料的统计学习,可以得到各词义与其上下文词语的搭配关系,从而获取汉语词义标注的知识.

统计学习的主要过程就是建立二维表及对二维表元素进行统计和累加. 二维表是 $M \times N$ 的矩阵,行 M 表示词义总数,每行表示一个语义群,对应《同义词词林》中的一个原子词群. 列 N 为语料中词语的总个数. 为了计算方便,单独用一个长度为 M 的向量存储各原子词群的编码,用一个长度为 N 的向量存储语料中的词语. 本文使用的语料库是搜狗实验室的中文词语搭配库(SogouR)[12]和抓取的短文本训练集. 首先从短文本训练集中抽取搭配关系按搭配库数据格式合并到 SogouR 中,SogouR 的数据格式为:

二元组 1　同现次数 1

二元组 2　同现次数 2

…

再按以下几下步骤进行词义标注的学习：

a）由同义词词林生成多义词典；

b）由多义词词典生成编码向量；

c）由词语搭配库生成搭配词向量；

d）生成义项判别矩阵.

为了在义项判别时搜索的高效性，对各个步骤生成的数据结构进行排序及建立索引. 词义义项判别学习的流程见图 2.

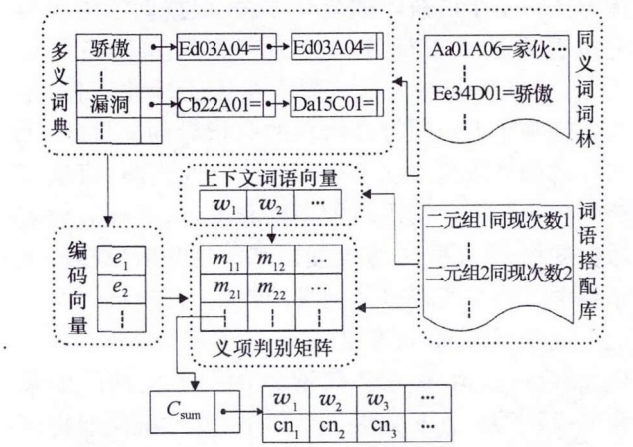

图 2　义项判别处理流程

Fig. 2　Flowchart of word meaning discrimination processing

图 2 中的义项判别矩阵具有两方面的功能，一是为多义词义项判别提供统计数据，二是为同义词扩展时确定扩展对象提供数据支持. 因此，矩阵中的元素要分别记录原子词群中各词与语料搭配词共现次数及共现次数之和. 下面给出生成义项判别矩阵的具体算法描述：

算法：义项判别矩阵生成算法.

输入：同义词词典（TongyiDic），词语搭配库（SogouR），多义词典（PolyDic），编码向量（CodeVector），搭配词向量（WordVector）.

输出：义项判别矩阵（Matrix）.

```
for each R_j in SogouR   //遍历每对搭配关系词
 {   //用 A 表示搭配关系 R_j 的左词，B 表示右词，Count 表示搭配关系出现的频率.
  if(PolyDic . find(A))   //若 A 为多义词
   {P=GetHead(A)   //获取 A 编码链首地址
    if(WordVector. find(B))   //若 B 为上下文词语向量中元素
     while(P++)
      {i=GetIndex(CodeVector, *P)   //获取当前编码在编码向量中的索引
       j=GetIndex(WordVector, B)   //获取 B 在上下文向量中的索引
       if(PolyDic. find(B))   //若 B 为多义词
        Matrix[i][j]+=count/H   //H 表示 B 的义项数，假定各义项在语料中均匀分布
       else
        Matrix[i][j]+=count
      }
   }
  if(PolyDic . find(B))   //若 B 为多义词
   {P=GetHead(B)   //获取 B 编码链表首地址
    if(WordVector. find(A))   //若 A 为上下文词语向量中元素
     while(P++)
      {   //获取当前编码在编码向量中的索引
       i=GetIndex(CodeVector, *P)
       //获取 B 在上下文向量中的索引
       j=GetIndex(WordVector, A)
       if(PolyDic. find(A))   //若 A 为多义词
        Matrix[i][j]+=count/L   //L 表示 A 的义项数，假定各义项在语料中均匀分布
       else
        Matrix[i][j]+=count
      }}
 }
```

义项判别矩阵生成后，利用下列公式近似估算先验概率 $p(s_i)$ 和条件概率 $p(v_k|s_i)$：

$$p(s_i)=\frac{c(s_i)}{\sum_{i=1}^{m}c(s_i)} \tag{2}$$

$$p(v_k\mid s_i)=\frac{c(s_i,v_k)}{c(s_i)} \tag{3}$$

其中：$c(s_i)$ 表示词义出现的次数，$c(s_i,v_k)$ 表示 s_i 与其上下文 v_k 共现的次数.

2　基于同义词词林扩展的短文本分类

在类似的上下文中交替使用同义词是常见的语言现象. 表达同一主题，不同的人用词也有各自的偏好，往往使用不同的词. 同义词的存在和使用丰富了文本的表达. 由于训练集的有限性，往往造成训练过程的特征维度缺失，在分类时，分类器将会一定程序上忽视缺失特征的同义表达，若事先通过同义词扩展补充特征信息，将对分类过程具有重要指导意义. 利用同义词对短文本进行扩展的基本过程是：对短

文本进行预处理之后剩下的原始词,从同义词词林中搜索包含该词的原子词群,将原子词群中的词作为候选扩展特征,加入到原始文本的特征向量中. 由于各原子词群中的词数并不相同,词数从两个到几十不等,而且,仅机械地将小类中的词扩展到原始文本中,反而会对分类产生不利影响. 原子词群中的哪些词能作为扩展特征?不仅要考虑与特征项的同义关系,而且要考虑在给定上下文背景下的语义关联. 为此,定义了语义关联强度函数 g,通过该函数指示候选扩展词是否最终作为扩展特征,基本思想是以原子词群各词与上下文在语料中的共现率为测度,以待扩展特征项与上下文在语料中的共现率为基准,通过设置阈值确定同义词是否作为扩展特征. g 函数计算步骤为:

1) 计算待扩展主干词 s_i 与上下文 $V_{\text{context}}=(v_1, v_2, \cdots, v_n)$(上下文长度为前后 7 个词)的共现概率 α:

$$\alpha = p(s_i, V_{\text{context}}) = \prod_{v_k \in V} p(v_k \mid s_i) \tag{4}$$

2) 分别计算主干词 s_i 的各个同义词 c_j 与上下文 $V_{\text{context}}=(v_1, v_2, \cdots, v_n)$的共现概率 β:

$$\beta = p(c_j, V_{\text{context}}) = \prod_{v_k \in V} p(v_k \mid c_j) \tag{5}$$

3) g 函数通过下式计算:

$$g(c_j) = \begin{cases} 1 & (\beta \geqslant \alpha) \\ 1 & (\beta < \alpha, \mid \beta - \alpha \mid < \varepsilon) \\ 0 & (\beta < \alpha, \mid \beta - \alpha \mid \geqslant \varepsilon) \end{cases}$$

g 结果为 1 时表示 c_j 为扩展特征,结果为 0 表示 c_j 为非扩展特征.

当 $\beta \geqslant \alpha$ 时,同义词 c_j 与上下文的联合概率比主干词 s_i 与上下文的联合概率大,说明 c_j 与上下文的关联性更强,c_j 自然确定为扩展特征.

当 $\beta < \alpha$ 时,同义词 c_j 与上下文的联合概率比主干词 s_i 与上下文的联合概率小,当两个概率值不小于设置阈值时,c_j 确定为扩展特征,当两个概率值小于设置阈值时,c_j 为非扩展特征. 特别地,当 $\beta=0$ 时,c_j 为非扩展特征.

同义词扩展算法具体描述如下:

步骤 1:对训练集和测试集中的每篇短文本 t 进行分词、去除停用词等预处理,得到 t 的主干词向量 $(w_1, w_2, \cdots, w_m)$.

步骤 2:对主干词向量中的每个词 w_i,在多义词词典中搜索,若 w_i 是多义词,根据式(1)计算各义项的概率值,将具有最大概率值义项的同义词列表作为候选扩展特征,若 w_i 为单义词,在同义词词典中搜索包含该词的原子词群,将原子词群中的词作为候选扩展特征.

步骤 3:计算候选扩展特征的语义关联强度 g 函数,择取 g 函数值等于 1 的候选扩展特征加入到原始特征向量中.

步骤 4:合并扩展后的重复特征. 由于特征之间有着丰富的语义联系,可能会出现扩展特征同原始特征词相同,或者扩展特征词有重复现象,这时需要合并相同的特征词,合并时将它们的权重相加求和.

步骤 5:使用传统文本分类算法进行训练学习和分类测试.

3　实验结果及分析

由于针对短文本分类的研究目前尚无公开的数据集,实验所用短文本语料由项目组从新浪、腾讯等各大网站抓取不同类别跟帖中的评论,包括军事、财经、体育、娱乐、游戏和科技 6 个类别,平均每类 8 000个文本. 分类实验中采用了 3 份交差验证,首先把整个语料分成 3 份,然后取其中的两份进行训练,剩余一份作测试,再把这 3 次的平均结果作为实验结果.

实验中分词处理采用中国科学院计算技术研究所的 ICTCLAS 系统,词语搭配库使用了搜狗实验室提供的中文词语搭配库(SogouR),同义词典使用了哈尔滨工业大学信息检索实验室的《同义词词林(扩展版)》. 分类方法选择朴素贝叶斯分类器,短文本特征表示使用 TF-IDF 法,特征选择使用 CHI. 分类器性能使用准确率 P、召回率 R 和 F_1 值进行评估. 共进行四组实验.

实验 1:无特征扩展的短文本分类,使用传统分类方法进行短文本分类,通过特征选择控制特征数进行多组测试,分类结果作为检验特征扩展算法的对比基准.

实验 2:不考虑多义词情况下基于同义词扩展的短文本分类,将每个短文本中的主干词进行同义词扩展,若主干词是多义词时,将主干词所有义项对应的同义词都进行扩展.

实验 3:考虑多义词判别情况下基于同义词扩展的短文本分类,将每个短文本中的主干词进行同义词扩展,若主干词是多义词时,进行义项判别,将判定的义项的同义词进行扩展.

实验 4:考虑多义词判别且进行候选扩展特征过滤的短文本分类,在实验③基础上,对候选扩展特征进行过滤,只对满足条件的候选特征进行扩展. 实验中的阈值 ε 通过设定不同取值观察分类结果确定.

实验结果见表1. 从表中可以看出,后两组实验

表 1 4 组实验结果

Tab. 1 Four groups of experimental result

类别	精确率				召回率				F_1 值			
	实验 1	实验 2	实验 3	实验 4	实验 1	实验 2	实验 3	实验 4	实验 1	实验 2	实验 3	实验 4
军事	70.45	69.88	74.76	77.14	73.56	72.90	78.26	80.42	71.97	71.36	76.47	78.75
财经	69.76	69.89	73.67	75.43	67.34	68.15	72.51	74.55	68.53	69.01	73.09	74.99
体育	72.74	73.43	75.92	78.47	68.90	67.75	73.47	76.68	70.77	70.48	74.67	77.56
娱乐	62.43	62.56	66.34	67.39	62.54	62.85	67.44	69.06	62.48	62.70	66.89	68.21
游戏	70.83	71.34	67.34	77.67	73.67	73.44	70.46	80.24	72.22	72.37	68.86	78.93
科技	60.86	60.25	65.55	66.56	63.68	63.75	67.32	68.07	62.24	61.95	66.42	67.31
均值	67.85	67.89	70.60	73.78	68.28	68.14	71.58	74.84	68.04	67.98	71.07	74.29

总体上优于前两组实验，无论准确率、召回率还是 F_1 值均有提高；第 2 组实验和第 1 组实验分类性能相当，在不同类别不同评估指标上彼此有所浮动；第 3 组实验在 5 个类别上分类精度高于实验 1，在第五个类别上略低于第 1 组实验，第 4 组实验较实验 3 评估指标又有了一定提高，准确率、召回率、F_1 值平均提高 3%左右，较实验 1 则平均提高 6%左右. 分析实验结果可以得出以下结论：

1）多义词对特征扩展有较大影响，未区别义项的特征扩展难有稳定的性能改善，有时甚至引入过多噪声降低分类性能，而在特征扩展时对多义词义项进行甄别有助于提高分类性能.

2）特征扩展时对多义词义项进行甄别基础上，是否对候选扩展特征进行过滤对提高分类性能也有较大影响，在特征扩展时会因为训练集本身的有效特征数目太少而又进行了过度扩展，反而造成分类错误；另一方面，结合给定上下文的语义关联性，对候选扩展特征进行过滤可以大幅提高分类性能，并且可以克服过度扩展带来的性能损失.

4 结论

本文提出了基于同义词词林扩展的短文本分类方法. 该方法利用词语之间的同义关系，通过引入外部语义资源《同义词词林》对原文本进行特征扩展，在扩展过程中对影响分类性能的几个问题进行了有效处理，利用大规模词语搭配库对多义词义项进行判别，通过计算待扩展短文本上下文与候选扩展特征的语义关联性实现对候选扩展特征的过滤. 初步实验结果显示，本文提出的方法可以有效提高短文本的分类效果.

参考文献：

[1] HOTHO A, STAAB S, STUMME G. Wordnet improves text document clustering [C]//Proceedings of the Semantic Web Workshop of SIGIR. Toronto: [s. n.], 2003: 541-544.

[2] STRUBE M, PONZETTO S P. Wikirelate ! Computing semantic relatedness using wikipedia [C]//Proceedings of the 21st National Conference on Artificial Intelligence. Boston: [s. n.], 2006: 1419-1424.

[3] GABRILOVICH V, MARKOVITCH S. Computing semantic relatedness using wikipedia-based explicit semantic analysis [C]//Proceedings of IJCAI. Hyderabad: [s. n.], 2007: 1606-1611.

[4] 王细薇，樊兴华，赵 军. 一种基于特征扩展的中文短文本分类方法 [J]. 计算机应用，2009，29(3)：843-845.

[5] 宁亚辉，樊兴华，吴 渝. 基于领域词语本体的短文本分类 [J]. 计算机科学，2009，36(3)：142-145.

[6] 王 盛，樊兴华，陈现麟. 利用上下位关系的中文短文本分类 [J]. 计算机应用，2010(3)：47-51.

[7] 范云杰，刘怀亮. 基于维基百科的中文短文本分类研究 [J]. 现代图书情报技术，2012(3)：47-52.

[8] 袁 满，欧阳元新，熊 璋，等. 一种基于频繁词集的短文本特征扩展方法 [J]. 东南大学学报：自然科学版，2014，44(2)：256-259.

[9] 同义词词林(扩展版) [EB/OL]. (2012-07-02) [2014-07-20]. http://www.datatang.com/data/42306/.

[10] 刘 挺，卢志茂，李 生. 一个全文词义自动标注系统的实现 [J]. 哈尔滨工业大学学报，2005，37(12)：1603-1605.

[11] 卢志茂，刘 挺，李 生. 基于无指导机器学习的全文词义自动标注方法 [J]. 自动化学报，2006，32(2)：228-236.

[12] 中文词语搭配库(SogouR) [EB/OL]. (2012-09-04)[2014-07-20]. http://www.sogou.com/labs/dl/r.html.