doi :10.3969/j.issn.1671-1122.2016.05.010

# 基于句义成分的短文本表示方法研究

尚海，罗森林，韩磊，张笈
（北京理工大学信息系统及安全对抗实验中心，北京 100081）

**摘 要：**随着移动互联网和信息技术的迅速发展，评论、微博等短文本数量呈现爆炸式增长。短文本数据少，文本特征稀疏，亟需有效的短文本表示方法来提升针对短文本的文本分类、文本聚类、热点发现、舆情分析等应用的效果。针对短文本特征稀疏问题，文章提出融合句义成分的短文本表示方法。该方法考虑短文本的语义信息，在保证特征空间维度不变的同时，结合句义成分和主题模型构建语义相关词语，再利用句义结构模型的话题和述题构建主题选择判定规则，选择语义相关词语扩充到短文本中，减少短文本表示向量中的0值特征。文章基于Sogou文本分类语料库，选择3个类别数据进行文本分类实验，并利用5折交叉方法选定模型参数。结果表明，文中方法对短文本分类的精确度达到0.7958，结果优于对比的短文本表示方法。因此，利用语义相关词语丰富短文本的语义信息，能够有效缓解短文本特征稀疏问题。文中短文本表示方法可以有效提高短文本分类等具体应用效果。

**关键词：**文本表示；句义成分；主题模型；文本分类
**中图分类号：**TP309 **文献标识码：**A **文章编号：**1671-1122（2016）05-0064-07

中文引用格式：尚海，罗森林，韩磊，等. 基于句义成分的短文本表示方法研究[J]. 信息网络安全，2016(5)：64-70.

英文引用格式：SHANG Hai, LUO Senlin, HAN Lei, et al. Research on Short Text Representation Based on Sentential Semantic Components[J]. Netinfo Security, 2016 (5) : 64-70.

# Research on Short Text Representation Based on Sentential Semantic Components

SHANG Hai, LUO Senlin, HAN Lei, ZHANG Ji
(*Information System and Security & Countermeasures Experimental Center, Beijing Institute of Technology, Beijing* 100081, *China*)

**Abstract:** With the development of mobile Internet and information technology, short text data such as commentary, microblog, has explosive growth. The sparseness of short text requires an effective algorithm of short text representation to improve the results of text clustering and classification, hot event detection and public opinion analysis, etc. This paper proposes an algorithm of short text representation based on sentential semantic components. Without changing the dimension of feature space, the method utilizes the sentential semantic components and topic model to obtain the semantic correlated words, and expands the short text with those words according to the topic selection rules. It reduces the zero-value dimension of in the text representation feature vectors. This paper implements short text classification experiments based on the Sogou corpus. The results show that the accuracy of short text classification reaches 0.7958, which is better than other methods. In summary, the proposed short text representation method, expanding short text with the semantic correlated words, can mitigate the sparseness problem effectively and improve the performance of short text classification.

**Key words:** text representation; sentential semantic components; topic model; text classification

**收稿日期：**2016-03-14
**基金项目：**国家242信息安全计划[2005C48]
**作者简介：**尚海（1990—），男，江苏，硕士研究生，主要研究方向为文本安全；罗森林（1968—），男，河北，教授，博士，主要研究方向为信息安全、数据挖掘、文本安全；韩磊（1985—），男，河北，博士研究生，主要研究方向为自然语言处理；张笈（1968—），男，陕西，副教授，博士，主要研究方向为信息安全。
**通信作者：**张笈 zhangji_bit@163.com

## 0 引言

文本表示方法属于自然语言处理中的基础性研究，在众多的应用中都需要对文本进行表示，如文本分类、文本聚类、自动摘要和舆情分析等。但随着移动互联网时代的到来，微博、评论、微信等短文本数据呈现爆炸式增长，对文本的处理提出了更高的要求。由于短文本内容短，因此使用传统的文本表示方法对短文本进行表示时，会存在特征稀疏问题。

目前，短文本的表示方法主要是基于向量模型，在特征表示上进行研究，通过改善特征表示方法解决特征稀疏问题。文献 [1] 利用搜索引擎计算词语之间的相似度，将相似度值用于文本相似度计算。但是由于需要使用搜索引擎，效率不高，同时对搜索结果中有效信息的抽取也是一个很困难的过程。文献 [2] 将短文本中的词语映射到 HowNet 知识库，通过知识库中的知识概念来丰富短文本的特征。但知识库的建立大部分是基于领域相关的，这给短文本表示带来了限制；引入知识概念对短文本进行表示的同时，也增加了特征空间的维度，提高了计算复杂度。还有一些学者利用数据统计的方法解决特征稀疏问题，文献 [3] 通过改进的 *TF*IDF* 算法表示 Twitter 微博文本。随着词向量的提出，文献 [4] 提出段落向量（Paragraph Vector，P2V）的短文本分布式表示方法，非监督地训练神经网络，将一句话、一个段落或一篇文本表示成一个向量。但单纯依靠统计方法，并没有利用到语言中的语义信息，忽略了词语在语义表达中的语义功能和词语之间的联系信息。

本文针对短文本表示特征稀疏问题，提出融合句义成分的短文本表示方法。该方法将句义结构模型的语义信息与主题模型结合，使用语义相关词语对短文本进行扩充，在保证特征空间维度不变的前提下，丰富短文本内容。

## 1 句义结构模型

为了表示中文句子的语义，展示句义的结构，根据贾彦德先生著作《汉语语义学》中的句义理论 [5]，罗森林 [6] 等人构建了汉语句义结构模型（Chinese Sentential Semantic Model，CSM）。通过该模型可以得到句子中每个词语在句义表达时所承担的语义功能，以及这些语义功能之间的关系 [7]。句义结构模型中，语义功能被称作句义成分，语义功能之间的关系被称作句义成分间关系。根据句义理论，句子的语义如公式（1）所示。

$$M=F(U,S) \quad (1)$$

其中，$M$ 表示句义，$F$ 表示句义理论中的规则，$U$ 表示义位（词义），$S$ 表示句义结构模型。句义结构模型是一种语义表示模型，对一条句子进行句义结构模型表示时得到该条句子的句义结构（Sentential Semantic Structure，SSS）。句义结构是句义结构模型针对具体句子的实例化结果，其定义如公式（2）所示。

$$s=\psi\left(\sum_{i=1}^{N}\sum_{j=1,j\neq i}^{N}\left(C_i,R_{ij}\right)\right) \quad (2)$$

其中，$s$ 表示句义结构，是句义的结构化表示；$\psi$ 表示句义成分和句义成分间关系的结合规则；$N$ 表示句子中句义成分的数量；$C_i$ 表示第 $i$ 个句义成分；$R_{ij}$ 表示句义成分 $C_i$ 和 $C_j$ 之间的关系。

### 1.1 句义成分

句义结构模型以《汉语语义学》[5] 为基础，从句义角度研究了句子的句义成分以及句义成分间关系的结构化表示模型，将抽象的句义表示成计算机可处理的结构化数据。如表 1 所示，模型包含的句义成分有句义类型、话题、述题、谓词、语义格和项。句义成分中的项分为基本项与一般项，其语义功能用语义格表示，对应的语义格分为 7 个基本格和 12 个一般格 [6]。如表 2 所示，基本格主要包括施事格、受事格、与格等；一般格主要包括范围格、时间格、描写格等。

**表1 句义成分类型及说明**

| 类型 | 说明 |
|---|---|
| 句义类型 | 句义结构的类别描述，反映句义结构的复杂程度、层次数目 |
| 话题 | 句义描述的对象，如句子“我吃苹果”中的“我” |
| 述题 | 句义对话题的描述内容，如“我吃苹果”中的“吃苹果” |
| 谓词 | 对话题描述对象的变化、运动、行为等的说明 |
| 项 | 在句义结构模型中表示词语，分为一般项和基本项 |
| 语义格 | 在句义中的语义功能，对应项的类型分为基本格和一般格 |

### 1.2 句义结构的形式

句义结构模型将句义结构分为句型层、描述层、对象层和细节层 4 个层次。其基本形式如所图 1 示。句型层中包含句义类型；描述层中包含话题（Topic）和述题（Comment）；对象层中包含语义格、谓词和项；细节层中

表2 语义格类型说明

| 语义格 | | 说明 |
|---|---|---|
| 基本格 | 施事格 | 变化、动作、行为的发起者 |
| | 遭遇格 | 变化、动作、行为的非自主发起者 |
| | 受事格 | 变化、动作、行为的承受者或产生的结果 |
| | 主事格 | 谓词所描述性质、状态等的对象 |
| | 说明格 | 协同谓词对主事格进行说明 |
| | 结果格 | 表示谓词的行为、动作的一种结果 |
| | 与格 | 某些谓词表示的行为、动作的间接对象 |
| 一般格 | 范围格 | 谓词所表示的动作、行为等活动的领域范围 |
| | 时间格 | 行为、运动、情感、状况发生或持续的时间 |
| | 空间格 | 谓词所表示的行为出现、发生的处所 |
| | 工具格 | 实现谓词表示行为所用的器具、材料等 |
| | 方式格 | 实现谓词表示行为或项的方式、方法、途径 |
| | 根由格 | 运动、行为等的依据、原因与目的 |
| | 属格 | 项所表示对象的拥有者、归属者 |
| | 描写格 | 对项所表示的对象起修饰作用 |
| | 同位格 | 与其后面的项共同描写、说明同一对象 |
| | 基准格 | 测量或比较谓词所表示情况的起算标准 |
| | 否定格 | 句义中表示否定 |
| | 其他格 | 不隶属于上述任一格的项 |

包含句子的引申含义。以汉语句义类型为准则，句义结构基本形式可以实例化为简单句义、复杂句义、复合句义、多重句义4种类型。

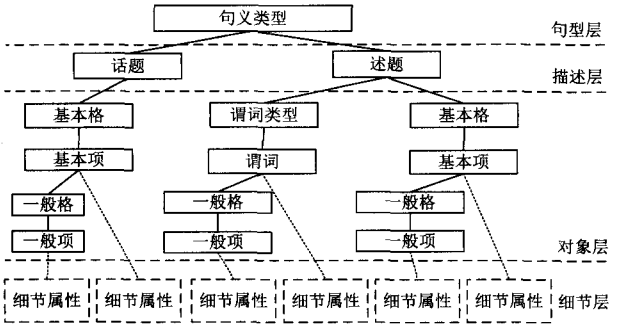



图1 句义结构的基本形式

细节层中的细节属性没有包含在句义成分中，因为细节属性是句义的引申含义或者附加含义。如果一般项是时间，如“1998年”，说明了句子描述内容所发生的时间，则这个项的语义格定义为时间格，同时可以引申出时间的细节属性；如果项中存在地点信息，如“北京”，说明了句子描述内容所发生的空间，则这个项的语义格定义为空间格，同时可以引申出空间的细节属性。目前，在句义结构模型中，细节层定义了谓词时态信息，并将其归入到了细节属性中。

## 2 短文本表示方法

融合句义成分的短文本表示方法核心思想是通过将语义相关词语填充到短文本中，扩充短文本中词语数量，解决短文本的特征稀疏问题。该方法包括两个步骤：1）得到语义相关词语；2）将语义相关词语填充到短文本中（特征丰富过程）。步骤1）利用句义结构模型中的句义成分，提取出文本集中基本项和一般项词语，使用主题模型对这些词语进行分析，将作为基本项和一般项的词语分配到不同的主题下，提取每个主题的*topN*（按照分布概率值排序，最大的前*N*个词语）个词语，认为这些词语即为语义相关词语。步骤2）同样利用句义结构模型中的句义成分，将文本中的基本项和一般项按照话题和述题进行划分，通过规则在主题中匹配这些基本项和一般项，选择一个主题下的*topN*个词语填充到短文本中。融合句义成分的短文本表示方法过程如图2所示。

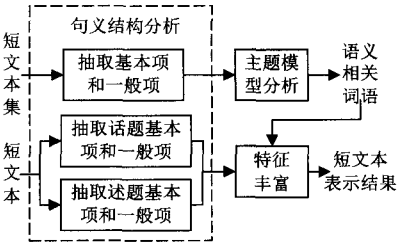



图2 融合句义成分的短文本表示方法过程

由图2可知，该方法的输入由短文本集和短文本两部分组成。输入短文本集是为了得到语义相关词语，输入短文本则是使用语义相关词语对其进行扩充，得到该短文本表示结果。对于同一批文本集，只需要进行一次分析即可得到语义相关词语。

### 2.1 句义结构分析

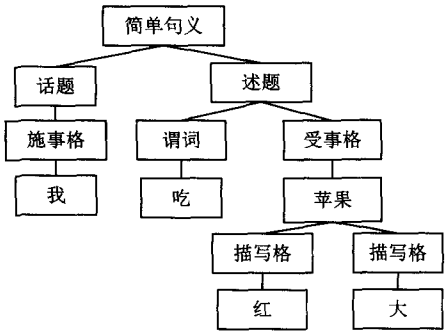
句义结构分析模块包括了段落分句、句义结构分析和去除停用词3个步骤。

当输入文本时，将文本进行分句处理，对每一条句子使用ICTCLAS 2015（http://ictclas.nlpir.org/）进行分词，利用句义结构分析方法[8]构建句义结构模型。输出是文本中的词语集合。

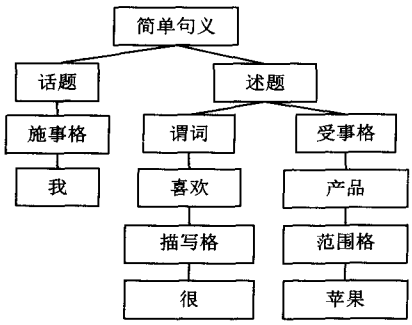
当输入文本集时，抽取文本中作为基本项和一般项的词语，输出用矩阵表示，矩阵的行对应每一篇文本，矩阵的列对应基本项或一般项的词语。

当输入单一短文本时，分别针对文本话题和述题抽取其中作为基本项和一般项的词语。同样，话题和述题的抽取结果分别用一个矩阵进行表示，矩阵的行对应每一篇短文本，矩阵的列对应话题或述题下的词语。

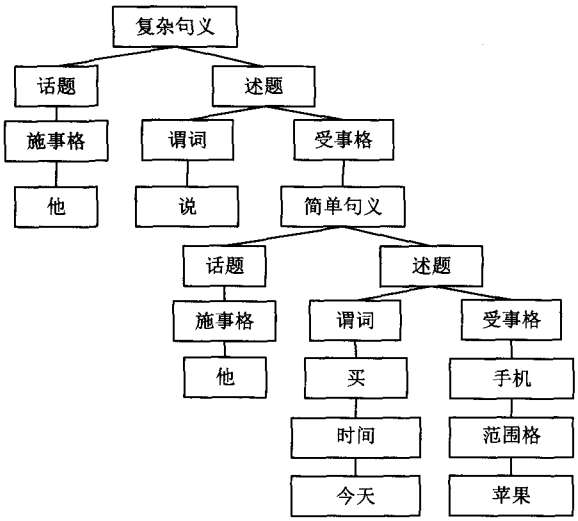
例如，短文本1“我吃了红色的大苹果”，短文本2“我很喜欢苹果的产品”和短文本3“他说他今天买了苹果手机”，它们的句义结构如图3所示。




a）短文本1句义结构图




b）短文本2句义结构图




c）短文本3句义结构图

**图3 短文本句义结构图**

当作为文本集输入时，短文本1对应输出文本中基本项词语集合“我”和“苹果”以及一般项词语集合“红”和“大”。短文本2对应输出文本中基本项词语集合“我”和“产品”以及一般项词语集合“很”和“苹果”。短文本3是复杂句义，可以理解为多个简单句义，输出文本中的“他”、“今天”、“苹果”和“手机”。

当作为短文本输入时，短文本1对应输出文本中话题下词语集合“我”和述题下词语集合“吃”、“苹果”、“红”和“大”。短文本2对应输出文本中话题下词语集合“我”和述题下词语集合“喜欢”、“产品”、“很”和“苹果”。短文本3输出“他”、“说”、“今天”、“买”“苹果”和“手机”。

### 2.2 主题模型分析

主题模型分析是为了得到主题下词语的分布情况，输入是文本集中基本项或一般项词语的集合，剔除了文本中的停用词、非语义词和谓词，对应输出是文本中主题下基本项和一般项的分布[9]。后续模块使用这些词语分布信息来扩充短文本内容，丰富短文本的特征。

本文实验采用LDA（Latent Dirichlet Allocation）模型[10]进行主题模型分析。LDA的概率图模型如图4所示。

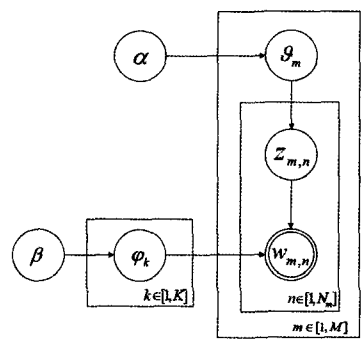



**图4 LDA概率图模型**

图4中，$\alpha$和$\beta$是狄利克莱超参数，$\vartheta_m$是文本$m$下主题的分布，$z_{m,n}$是第$m$篇文本的第$n$个词语$w_{m,n}$的主题，$\varphi_k$是第$k$个主题下词语的分布，$K$表示主题总个数，$N_m$表示第$m$篇文本中总词数（不去重），$M$表示数据集中总文本个数，单圆形表示隐藏变量，双圆形表示可观测变量，方框表示重复操作。

LDA模型基于文本独立性假设和词语独立性假设（词袋假设）。文本独立性假设认为文本之间相互独立不存在联系；词袋假设认为词语之间相互独立，将文本看成是词语的集合，不考虑词语之间的关系。LDA模型文本集生成过程如下：

1）对每一个主题$k$生成“主题-词语”分布$\varphi_k \sim Dir(\beta)$。

2）对每一篇文本$d$生成文本主题分布$\vartheta_m \sim Dir(\alpha)$；对该文本中第$n$个词语生成主题项$z_{m,n}$，生成词项$w_{m,n}$。

LDA模型针对第$m$篇文本的生成概率如公式（3）所示。

$$p(w|\alpha,\beta)=\prod_{n=1}^{N_d}p(w_{m,n}|\varphi_{z_{m,n}})p(z_{m,n}|\vartheta_m)p(\vartheta_m|\alpha)\cdots\cdots(3)$$

LDA主题模型原本的输入是文本中所有词语（去除停用词），在这里则剔除了文本中的非语义词和谓词。根据句义结构模型理论，非语义词在句义表达时没有作用，这种词语没有真实的语义，所以被剔除。谓词作为语义表达的核心内容，其语义搭配形式比基本项和一般项更加灵活，在句义结构模型中，谓词既与话题中的基本项结合，表示话题基本项的动作、行为和状态等，又与述题中的基本项结合，表示话题中基本项的结果、过程和受事等。例如，可以作为基本项或一般项的词语“医院”，一般情况下都会出现在与医院相关的段落或文本中，但是谓词“吃”既可以出现在描述人吃食物的段落中，又能出现在描述动物吃食物的段落，还能出现在其他领域，如“IBM被联想吃掉了”。所以谓词也被排除掉，只用基本项和一般项进行分析，得到它们在主题下的分布，减少谓词在后续短文本内容扩充时造成的影响。

### 2.3 特征丰富

特征丰富是将主题模型分析得到的主题下的词语添加到短文本中，丰富短文本内容，减少短文本因数据少而造成的特征稀疏问题。其输入包括短文本中话题下的基本项和一般项、述题下的基本项和一般项、主题模型分析结果中主题下词语的分布，输出短文本的文本向量。

特征丰富模块处理过程是：根据主题选取规则在LDA输出的主题词语分布中查询得到一个主题，将该主题下的*topN*词语填充到短文本中，丰富短文本中词语内容。主题选取规则是：根据文本中的词语，选择词语概率值和最大的主题。概率值和分为4种：$P^i_{tb}$话题下所有基本项在主题$i$下的概率值和；$P^i_{cb}$述题下所有基本项在主题$i$下的概率值和；$P^i_{tc}$话题下所有一般项在主题$i$下的概率值和；$P^i_{cc}$述题下所有一般项在主题$i$下的概率值和。计算方法如公式（4）所示。

$$P^i_*=\sum_{n=1}^{N}P^i_n\cdots\cdots(4)$$

其中，$P^i_*$表示*在主题$i$下的概率值和；$N$表示*中的不重复词语个数；*表示上述话题和述题的基本项和一般项；$P^i_n$表示词语$n$在主题$i$下的概率值，主题$i$下存在词语$n$则取对应概率值，不存在则取为0。通过概率值和选择词语的主题，选择判定规则如下：

$N_T$：LDA输入主题的个数

$P[N_T]$：存放$P^i_*$计算结果的临时数组

$L[N_T]$：与$P[N_T]$对应存放主题编号

For (int $i=0$; $i<N_T$; $i$ ++)

根据公式（4）计算$P^i_{tb}$、$P^i_{cb}$、$P^i_{tc}$和$P^i_{cc}$；

对$P^i_{tb}$按照数值由大到小排序存入$P[N_T]$，对应主题编号存入$L[N_T]$；

If ($P[0]>P[1]$)

return $L[0]$（被选主题的编号）；

else

对$P^i_{cb}$按照数值由大到小排序存入$P[N_T]$，对应主题编号存入$L[N_T]$；

If ($P[0]>P[1]$)

return $L[0]$（被选主题的编号）；

else

对$P^i_{tc}$按照数值由大到小排序存入$P[N_T]$，对应主题编号存入$L[N_T]$；

If ($P[0]>P[1]$)

return $L[0]$（被选主题的编号）；

else

对$P^i_{cc}$按照数值由大到小排序存入$P[N_T]$，对应主题编号存入$L[N_T]$；

return $L[0]$;

通过上述判定规则，得到语义相关词语的主题编号，使用该编号主题下的*topN*个词语对短文本进行扩充。词语填充完毕之后，使用短文本中所有词语作为特征构建向量，具体方法为：参照文献[3]中的方法计算词语的*TF*IDF*值，填充词语的维度值设置为1，输出向量即为短文本表示的结果。

## 3 短文本表示方法对比实验

### 3.1 实验数据源

数据源为Sogou文本分类语料库，来源于Sohu新闻网站保存的大量经过编辑整理与分类的新闻语料和对应的分类信息，有汽车、财经、IT、健康、体育、旅游、教育、招聘、文化和军事10个类别，约为十万篇文本。从数据源中随机选取汽车、财经、IT的600篇篇幅较短（句子个数小于等

于4，单个句子词数小于等于60）的文本作为分类实验原始数据，分布情况如表3所示。

**表3 短文本实验数据分布情况**

| 类别 | 汽车 | 财经 | IT | 总计 |
|---|---|---|---|---|
| 数目/篇 | 200 | 200 | 200 | 600 |

实验数据去除了原始数据中的乱码、空格等非正常字符，使用NLPIR 2015分词去除停用词，词性标记依据NLPIR 2015中的北大二级标准，每篇文本的平均词数为50，文本中最多包含词数为125，最少包含词数为4，去除重复后总词数为9027。

### 3.2 实验目的

为验证短文本表示效果，将本文提出的短文本表示方法与现有主要短文本表示方法分别应用于文本分类任务，比较本文方法与其他短文本表示方法应用于文本分类的效果。

### 3.3 评价方法

本文使用准确率（$P$）、召回率（$R$）和F-Measure值（$F$）对实验结果进行评价，准确率、召回率和F-Measure值的计算方法如公式（5）、公式（6）和公式（7）所示。

$$P=C/T \quad (5)$$

$$R=C/L \quad (6)$$

$$F=2PR/(P+R) \quad (7)$$

其中，在计算准确率和召回率时，$C$为某一类别被正确分类的样本总数；$T$指被预测为该类别的样本总数；$L$为在数据集合中该类别样本总数。

使用精确度计算结果评价所有类别的整体效果，精确度计算方法如公式（8）所示。

$$Accuracy=\frac{\text{正确分类样本数量}}{\text{总样本数}} \quad (8)$$

其中，在计算精确度时，正确分类样本数量指所有类别被正确分类的样本数，总样本数是数据集中样本总数。

### 3.4 实验过程和结果

短文本表示方法对比实验过程如图5所示。本文选取了基于统计的方法，包括LDA、P2V[4]、改进的*TF*IDF*[3]和基于主题扩充的方法[9]作为对比方法进行实验。其中，LDA方法、P2V方法和改进的*TF*IDF*方法都属于基于数据统计的方法[10]；基于主题扩充的方法属于基于知识概念的方法。对比实验采用相同的数据，基于不同方法得到短文本表示结果，通过5折交叉，使用SVM[11]进行分类得到实验结果。

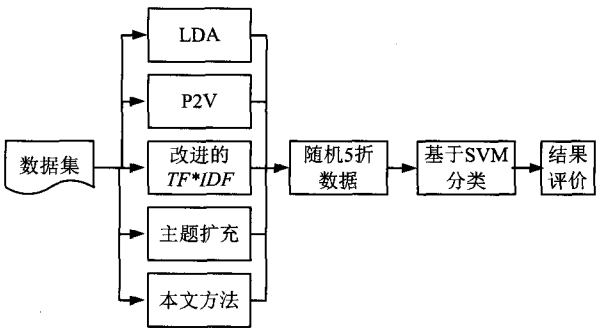



**图5 短文本表示方法对比分类实验流程**

实验首先对各类短文本表示方法基于5折交叉方法进行参数选择，以精确度值为标准，选择取得最优结果的模型参数[12]。LDA最终选择主题个数$K$=60，参数$\alpha$和$\beta$使用GibbsLDA++（http://gibbslda.sourceforge.net/）中的默认值，即$\alpha=50/K$（$K$为主题个数），$\beta$=0.01。P2V向量长度选择20，窗口长度选择9。基于主题扩充的方法仅调整LDA生成的主题个数，主题个数为10。改进的*TF*IDF*方法基于词频统计，不需要进行参数选择[13]。本文提出的融合句义成分的短文本表示方法中，主题模型使用了LDA模型。句义成分与LDA结合的短文本表示方法设置主题个数为20，每个主题下词语个数为40。

在以上得到各短文本表示方法最优参数的基础上，使用SVM算法对不同的短文本表示方法进行文本分类[14,15]，得到实验结果如表4所示。

**表4 短文本表示方法对比分类实验结果**

| | | LDA | 改进的*TF*IDF* | 主题扩充 | P2V | CSM+LDA |
|---|---|---|---|---|---|---|
| 汽车 | *P* | 0.9140 | 0.9615 | 0.9707 | 0.9474 | 0.9651 |
| | *R* | 0.7969 | 0.7031 | 0.7250 | 0.7875 | 0.7781 |
| | *F* | 0.8514 | 0.8123 | 0.8301 | 0.8601 | 0.8616 |
| 财经 | *P* | 0.9690 | 0.9227 | 0.9515 | 0.9028 | 0.9211 |
| | *R* | 0.3906 | 0.6344 | 0.6125 | 0.6094 | 0.6563 |
| | *F* | 0.5568 | 0.7519 | 0.7452 | 0.7276 | 0.7664 |
| IT | *P* | 0.5562 | 0.6047 | 0.6039 | 0.6444 | 0.6435 |
| | *R* | 0.9594 | 0.9563 | 0.9719 | 0.9625 | 0.9531 |
| | *F* | 0.7041 | 0.7409 | 0.7449 | 0.7719 | 0.7683 |
| 精确度 | | 0.7156 | 0.7646 | 0.7698 | 0.7865 | 0.7958 |

### 3.5 实验结果分析

由表4可知，句义成分与主题模型结合的方法（CSM+LDA）的分类精确度达到0.7958，结果优于P2V以及其他方法的结果。CSM+LDA方法与改进的*TF*IDF*方法相比，增加了扩充的词语，精确度比改进的*TF*IDF*提

高近 0.03；而与 LDA 方法相比，增加了句义结构分析，分类的精确度提高了 0.08。P2V 利用深度学习的方法分析文本的上下文信息，能够挖掘出文本的语义信息，得到较好的文本表示，结果优于其他对比方法，本文方法的分类精确度比 P2V 提高 0.01，能够挖掘深层次语义信息。

短文本表示方法对比实验的结果说明，在统计方法的基础上，加入语义信息的指导作用，有效提高了短文本表示效果，使得短文本表示结果在分类任务中提升了文本分类效果，也证明了本文提出的短文本表示方法的有效性。

## 4 结束语

本文针对短文本表示中存在的特征稀疏问题，提出了融合句义成分的短文本表示方法。该方法包括构建语义相关词语和短文本扩充两个步骤：构建语义相关词语按照词语的句义成分得到文本中作为基本项和一般项的词语，将这些词语使用 LDA 进行分析，得到“主题－词语”分布，每个主题下的词语即为语义相关词语；短文本扩充将短文本中词语分为基本项和一般项，再按照句义结构模型中的话题和述题对基本项和一般项进行第二次划分，得到话题下和述题下的基本项与一般项，使用主题选择规则将语义相关词语扩充到文本当中。

本文使用 Sogou 文本分类语料库中的数据进行了短文本表示方法对比实验，实验结果表明，本文的短文本表示方法分类精确度为 0.7958，相比 LDA、P2V、改进的 *TF*IDF* 等方法均有一定程度的提高。因此，融合句义成分的短文本表示方法通过对短文本内容的扩充，在不改变特征空间维度的前提下，减少了短文本的 0 值特征，有效缓解了短文本的特征稀疏问题，并能提升短文本的分类效果。（责编 潘海洋）

**参考文献：**

[1]WANG Meng, LIN Lanfen, WANG Jing, et al. Improving Short Text Classification Using Public Search Engines[C]//NAFOSTED, Springer. 2013 International Symposium on Integrated Uncertainty in Knowledge Modelling and Decision Making, July 12-14, 2013, Beijing, China. Heidelberg: Springer, 2013: 157-166.

[2]NING Yahui, ZHANG Li, JU Yarong, et al. Using Semantic Correlation of HowNet for Short Text Classification[EB/OL]. https://www.researchgate.net/publication/269359136_Using_Semantic_Correlation_of_HowNet_for_Short_Text_Classification, 2016-1-22.

[3]BEAUX B P. SHARIFI, INOUYE D I, KALITA J K. Summarization of Twitter Microblogs[J]. The Computer Journal, 2014, 57(3): 378-402.

[4]LE Q, MIKOLOV T. Distributed Representations of Sentences and Documents[C]//IMLS. 31st International Conference on Machine Learning, June 21-26, 2014, Beijing, China. Los Alamos: Eprint Arxiv, 2014: 1188-1196.

[5] 贾彦德．汉语语义学 [M]. 北京：北京大学出版社，1999.

[6] 罗森林，韩磊，潘丽敏，等．汉语句义结构模型及其验证 [J]. 北京理工大学学报，2013，33（2）：166-171.

[7] 冯扬．汉语句义模型构建及若干关键技术研究 [D]. 北京：北京理工大学，2010.

[8]LUO Senlin, HAN Lei, PAN Limin, et al. Construction Method of Chinese Sentential Semantic Structure[J]. Journal of Beijing Institute of Technology, 2015, 24(1): 110-117.

[9] 吴坚，沙晶．基于随机森林算法的网络舆情文本信息分类方法研究 [J]. 信息网络安全，2014（11）：36-40.

[10]BLEI D M, NG A Y, JORDAN A Y. Latent Dirichlet Allocation[J]. The Journal of Machine Learning Research, 2003, 3(1): 993-1022.

[11] 戚名钰，刘铭，傅彦铭．基于 PCA 的 SVM 网络入侵检测研究 [J]. 信息网络安全，2015（2）：15-18.

[12]VO D T, OCK C Y. Learning to Classify Short Text from Scientific Documents Using Topic Models with Various Types of Knowledge[J]. Expert Systems with Applications, 2015, 42(3): 1684-1698.

[13] 吴旭，郭芳毓，颉夏青，等．面向机构知识库结构化数据的文本相似度评价算法 [J]. 信息网络安全，2015（5）：16-20.

[14]CHANG C C, LIN C J. LIBSVM: A Library for Support Vector Machines[J]. ACM Transactions on Intelligent Systems and Technology, 2011, 2(3): 1-27.

[15] 高悦，王文贤，杨淑贤．一种基于狄利克雷过程混合模型的文本聚类算法 [J]. 信息网络安全，2015（11）：60-65.