一种基于词聚类的文本特征描述方法①

陈 炯[1]，张永奎[2,3]

[1](山西职业技术学院 计算机工程系，太原 030006)
[2](山西大学 计算机与信息技术学院，太原 030006)
[3](山西大学 计算智能与中文信息处理教育部重点实验室，太原 030006)

摘 要：针对文本挖掘中存在的特征空间高维性问题，提出了一种基于词聚类的文本特征描述方法，旨在通过机器学习的方法挖掘词汇之间的语义关联,动态构造特定领域的概念词典，借助构造的概念来描述文本的特征，该方法不借助主题词典,先从训练语料中对词的共现情况进行分析,用词聚类(word clustering)生成由种子词(seed words)表示的代表某一主题概念的词类，然后用种子词作为文本的特征项。实验表明，该方法不仅压缩了特征空间的维数，也克服了 HowNet 中概念信息的局限性,提高了文本分类的精确度。

关键词：文本特征描述；词共现；词聚类；种子词

A Description Method of Text Feature Based on Word Clustering

CHEN Jiong[1], ZHANG Yong-Kui[2,3]

[1](Department of Computer Engineering, Shanxi Polytechnic College, Taiyuan 030006, China)
[2](School of Computer & Information Technology, Shanxi University, Taiyuan 030006, China)
[3](Key Laboratory of Ministry of Education for Computation Intelligence and Chinese Information Processing, Shanxi University, Taiyuan 030006, China)

Abstract: Feature space has the high-dimensional problem in text mining. This paper presented a new description method of text feature based on word clustering. The purpose is to mine semantic association between words using machine learning, then to construct the concept dictionary in specific areas dynamically, finally to describe the text feature with the concept constructed. This method analyzes the co-occurrence of words in training corpus firstly, without using theme dictionary, then generates word cluster expressed in seed words which represents a concept of theme by word clustering, finally takes the seed words as text features. The experimental results indicate that this method not only reduces dimensionality of feature space but also overcomes the limitations of the concept in HowNet, and improve the performance of text categorization.

Keywords: text feature description; word co-occurrence; word clustering; seed words

1 引言

在文本分类、信息检索及信息过滤等文本挖掘应用中，为便于计算机处理无结构的文本，广泛采用向量空间模型(VSM)来描述文本，一个文本经过预处理后首先表示为 n 维特征空间中的一个 n 维向量，然后才对它进行相应的处理。特征空间的构造需要从训练文本集中抽取并选择具有代表性的特征项作为特征空间的维，因此，特征的选择和重构成为文本信息处理的基础环节[1]。

在实际应用中，特征空间的高维性加剧了机器学

① 基金项目:国家自然科学基金(60475022);山西省工业攻关项目(2006031178)
收稿时间:2010-06-18;收到修改稿时间:2010-08-03

习的负担，增加了计算的复杂度，降低了文本处理的精度。研究人员围绕特征空间的维数约减问题进行了大量的探索，并取得了一定的效果，这些探索可分为两类，一是通过构造各类评估函数，直接从原始特征中挑选出一些最具代表性的字、词、词组或短语作为特征[2]，如文档频率(DF)方法，信息增益(IG)方法，互信息(MI)方法,CHI 方法，期望交叉熵(ECE)，文本证据权(WET)，优势率(OddRatio)，基于词频覆盖度的特征选择方法等。但是，由于词语本身存在同义、多义以及对短语和上下文的依赖等现象，单纯基于词形的技术中，把意义可能密切相关的词孤立提取，忽略了词语的语言学特征和相互关系，因此导致这种特征提取存在较大的局限性。例如，传统的向量空间模型最基本的假设是各个分量之间正交，而实际上在真实文本中，作为分量的特征词往往有很大的相关性[3]。二是采用映射或变换的方法把原始特征变换为较少的新特征[2]，如主分量分析的方法，潜在语义索引等。清华大学廖莎莎，江铭虎等利用 HowNet 概念词典中的信息，抽取概念作为特征来构成文本向量[4]。由于概念空间比词空间小，而且各分量之间相对独立，因此，概念特征比词特征更适合用来表示文本内容。但是，HowNet 词典中的概念有限，不能够涵盖网上出现的大量新词，也无法跟踪网上词语的演化。基于此，我们提出基于词聚类的文本特征描述方法，旨在通过机器学习的方法动态构造特定领域的概念词典，借助构造的概念来描述文本的特征，该方法不借助主题词典,先从训练语料中对词的共现情况进行分析,用词聚类生成由种子词表示的代表某一主题概念的词类，然后用种子词作为文本的特征项。对于给定文本，先采用信息论方法进行关键词抽取，然后将关键词对应到特征空间，得到文本的特征向量描述。

2 词聚类

在文本信息处理中，概念的产生一般有三种方法[5]。一是用两个词相加得到的新词作为概念；二是将词的上位直接作为概念；三是通过若干意义相近的词聚类产生类中心，将类中心作为概念。本文采用第三种方法，用种子词作为类中心，把与某一主题相关联的词聚类，并用种子词表示该词类，每一个种子词代表一个主题概念。

词聚类这个术语过去已有文章使用[6,7]，但本文使用的聚类方法以及用途却与它们不同。从语义角度上讲，同义词或近义词是指在不同的场景下，不同的词表达的意义相同或相近，而概率意义下的同义词与近义词则具有更为广泛的含义，同义词或近义词指的是在某一主题下，它们的概率关联较为密切[1]。所谓种子词是指与某一主题密切相关并能够代表该近义词类的词。例如：爆炸（煤矿，炸弹，恐怖，自杀，汽车，瓦斯，死亡）是以种子词“爆炸”表示的一个词类，并称词类中的任意一个词为该词类或表示该词类的种子词的一个元素。如：“煤矿”、“炸弹”、“恐怖”、“自杀”等分别为种子词“爆炸”表示的词类的一个元素，或称为种子词“爆炸”的一个元素。

采用 CHI 统计法计算词 w_i 与 w_j 之间的相关度，因为这种方在处理中文文本时具有较好的性能[8]。假设 w_i 与 w_j 是两个词，且 w_i 与 w_j 之间符合具有一阶自由度的 χ^2 分布， w_i 对于 w_j 的 χ^2 统计值越大，表明它们之间的相关度越大。令 n 表示训练语料中文本总数，a 表示 w_i 出现且 w_j 也出现的文本频数，b 表示 w_i 出现但 w_j 不出现的文本频数，c 表示 w_j 出现但 w_i 不出现的文本频数，d 是 w_i 与 w_j 均不出现的文本频数，则 w_i 与 w_j 的相关度值由(1)式计算：

$$\chi^2(w_i,w_j)=\frac{n*(a*d-c*b)^2}{(a+c)*(b+d)*(a+b)*(c+d)} \quad (1)$$

在训练文本集中，采用如下算法获得用种子词表示的特征词：

1) 文本预处理。对训练文本集中的文本进行分词，去掉停用词、连词、代词、冠词等,得到候选词集 $U=\{w_1,w_2,...,w_n\}$。

2) 种子词选取。对U中的词进行词频统计，并选取频度大于阈值 f 的词构成种子词集 $V=\{sw_1,sw_2,...,sw_m\}$。

3) 词聚类。对于候选词集 U 中每一个词 $w_j(j=1,2,...,n)$,按(1)式依次计算它与种子词集 V 中的每一个种子词 $sw_i(i=1,2,...,m)$ 的相关度 $\chi^2(sw_i,w_j)$。若 $\chi^2(sw_i,w_j)\ge r$（r 是给定的聚类阈值），且满足 $\chi^2(sw_i,w_j)=\max\limits_{1\le q\le m}\{\chi^2(sw_q,w_j)\}$，则将词 w_j 归入种子词 sw_i 表示的词类，否则继续计算 U 中下一个词与种子词 $sw_i(i=1,2,...,m)$ 的相关度。直到 U 中所有词都计算完毕，得到由种子词表示的词类，记为 $sw_i(w_{i1},w_{i2},...,w_{ik})$，其中 sw_i 为第 i 个种子词，$w_{i1},w_{i2},...,w_{ik}$ 为种子词 sw_i 的元素，sw_i 为 $w_{i1},w_{i2},...,w_{ik}$ 所属的种子词。文本的特征可采用种子词向量 $(sw_1,sw_2,\cdots,sw_m)$ 来描述，这样可以获得文本中最重要且相互独立的用种子词表示的特征，在很大程度上降低了特征空间的维数。

4) 词类权重因子计算。设 $sw_i(w_{i1},w_{i2},...,w_{ik})$ 是由种子词 sw_i 表示的词类，以词类所含平均信息量作为该词类

权重因子，按(2)式计算：

$$G(sw_i)=-\frac{1}{k}\sum_{j=1}^{k}p(w_{ij})log\ p(w_{ij}) \quad (2)$$

其中，$G(sw_i)$表示种子词 sw_i 的权重因子，$p(w_{ij})$表示种子词 sw_i 的第 j 个元素的概率分布。词类权重因子反映了特征项在特征空间中区分文本类别属性的能力。

3 文本特征描述

给定文本 d，我们通过关键词抽取、特征生成和权重计算三个步骤来生成文本 d 的特征向量描述。

3.1 关键词抽取

先对给定文本 d 进行预处理(分词，停用词处理)，使用香农信息论对给定文本 d 进行关键词抽取[9]。经预处理后的文本可看作一个词序列，用 $d=(w_1,w_2,\cdots,w_n)$表示，假设文本 d 按照一个离散的概率分布 p(w)独立地生成，其中随机变量 w 在词汇集中取值，根据信息论，用 $H(w_i)$ 表示词 w_i 在文本中所含信息量，计算公式如下：

$$H(w_i)=-N(w_i)\times log\ p(w_i) \quad (3)$$

其中，$N(w_i)$表示词在文本 d 中出现的频次，$p(w_i)$为词 w_i 的概率分布，可采用极大似然估计方法计算，计算方法为：$p(w_i)=F(w_i)\times F$，其中 F 表示训练文本集中总词频数，$F(w_i)$表示训练文本中词 w_i 出现的频次。将文本中所有词按 $H(w_i)$ 值降序排列，选取 $H(w_i)$大于某个阈值的词作为该文本关键词。

3.2 特征生成

给定文本 d，经关键词抽取产生 h 个关键词 $w_1,w_2,\ldots,w_h$，采用以下方法实现关键词到特征空间的映射：

(1) 若文本 d 的第 i 个关键词 w_i $(1\le i\le h)$是种子词，则直接将其映射为文本 d 的特征词；

(2) 若文本 d 的第 i 个关键词 w_i $(1\le i\le h)$不是种子词，且 w_i 是种子词 sw_i 的元素，则将文本 d 的关键词 w_i 映射为由种子词 sw_i 表示的特征词；

(3) 若文本 d 的第 i 个关键词 w_i $(1\le i\le h)$不是种子词，且 w_i 无所属种子词，则将关键词 w_i 从 d 的特征中去除。

3.3 权重计算

根据文本 d 的关键词到特征空间的映射方式的不同，计算文本在各个特征上的权重值。

若文本 d 的第 i 个关键词 w_i $(1\le i\le h)$是特征词，则特征词的权重按式(4)计算。

$$Q(sw_i)=G(sw_i)H(w_i) \quad (4)$$

其中，$Q(w_i)$为特征词 w_i 的权重，$G(w_i)$为词类权重因子，$H(w_i)$为词 w_i 在文本 d 中所含信息量。

若文本 d 的第 i 个关键词 w_i $(1\le i\le h)$不是种子词，且 w_i 是种子词 sw_i 的元素，则将种子词 sw_i 作为文本 d 特征词,权重按式(5)计算。

$$Q(sw_i)=\chi^2(sw_i,w_j)G(sw_i)H(w_i) \quad (5)$$

其中，$Q(w_i)$为关键词 w_i 对应的特征权重，$\chi^2(w_i,w_j)$为词 w_i 与 w_j 的相关度，$G(w_i)$为词类权重因子，$H(w_i)$为词 w_i 在文本 d 中所含信息量。

然而，在特征生成中可能会存在两个或多个关键词生成的特征词相冲突的情况，也就是说，若 d 有两个关键词 w_i 与 w_j，其中 w_i 为种子词，w_j 不是种子词，但 w_j 是种子词 w_i 的元素，按照上述特征描述生成方法，这两个词生成的特征词都是 w_i，则将这两个关键词所生成的特征词合并，合并后的特征词权重按式(6)计算。

$$Q(w_i)=(1+\chi^2(w_i,w_j))G(w_i)H(w_i) \quad (6)$$

其中 $Q(w_i)$为合并后的特征词权重，$\chi^2(w_i,w_j)$为词 w_i 与 w_j 的相关度，$G(w_i)$为词类权重因子，$H(w_i)$为词 w_i 在文本 d 中所含信息量。

为方便说明，假设在文本训练过程中只产生了三个词类，它们分别是：煤矿 0.18（煤矿，矿难，事故，瓦斯）；爆炸 0.21（爆炸，炸弹，恐怖，自杀，汽车，瓦斯，死亡），事故 0.13（事故，矿工，抢救）括号前的“煤矿”、“爆炸”和“事故”都是种子词，种子词后的数字为种子词所表示的词类的权重因子，这三个种子词将作为文本的三个特征词，文本的特征空间描述为（(煤矿，Q_1)，(爆炸，Q_2)，(事故，Q_3)）。又假设给定文本 d 经关键词抽取后只有五个关键词为：煤矿(0.63)；爆炸(0.77)，死亡（0.43），矿工（0.27），遇难（0.21），其中括号中的数字分别为关键词在文本 d 中所含信息量。则按照本文方法，由于给定文本 d 的五个关键词中只有两个词“煤矿”与“爆炸”是种子词，另外三个词“死亡”、“矿工”和“遇难”不是种子词，但关键词“死亡”可对应到种子词“爆炸”上，应与关键词“爆炸”合并计算权重，关键词“矿工”可对应到种子词“事故”上，应与关键词“事故”合并计算权重，而关键词“遇难”无所属的种子词，则将其去除。设关键词“爆炸”与“死亡”的相关度为 0.62，关键词“事故”与“矿工”的相关度为 0.33，

则所生成的文本 d 的向量描述为：d=((煤矿，0.1134)，(爆炸，0.2620)，(事故，0.0116))。

4 实验设置及结果分析

4.1 实验设置

为了验证本文所提出的基于词聚类的文本特征描述方法的实际性能，我们在国家语委现代汉语语料库的 5 类文档集中各随机抽取不重复的 60 篇文本，共 300 篇文本，形成一个文档集。采用同样的方式共形成 5 个文档集,针对上述 5 个文档集，我们进行了三项实验。

实验中为了选取合适的阈值参数，首先对训练文本进行分词处理，去掉连词、代词、冠词及低频词等；根据对文本特征的打分情况设置初始阈值，然后利用本文方法构造文本特征集，并对训练文本进行特征加权表示；通过 KNN 分类学习算法对训练文本进行类别学习，并用训练获得的分类器对训练文本进行封闭测试。如果评价结果达到最优则停止训练，得到最终分类器；如果评价结果没有达到最优，则根据评价结果对阈值参数进行调整，重新构造特征集，并返回继续进行训练。通过上述方法，实验选取的阈值参数为：种子词选取中的频度阈值 f=18，词聚类中的相关度阈值 r=0.005。

4.2 评测标准

采用特征压缩比 TP 来评测基于词聚类的文本特征描述方法的维数约减效果，特征压缩比 TP 采用公式(7)计算。

$$TP=\frac{\text{原始特征词数}-\text{结果特征词数}}{\text{原始特征词数}}\times 100\% \tag{7}$$

采用宏平均准确率 MP、宏平均召回率 MR 和宏平均 F_1 值 MF_1 评测本文方法对文本分类算法性能的影响，MP，MR 和 MF_1 的计算公式如下：

$$MP=\frac{1}{n}\sum_{i=1}^{n}P_i \tag{8}$$

$$MR=\frac{1}{n}\sum_{i=1}^{n}R_i \tag{9}$$

$$MF1=\frac{2\times MP\times MR}{MP+MR} \tag{10}$$

其中 P_i 为第 i 类的准确率，R_i 为第 i 类的召回率，n 为训练集分类数。

4.3 实验结果及分析

实验一：本文方法对特征维数约减的效果测试

为了验证本文方法的特征维数约减效果，首先将 5 个文档集中的文本经过预处理，去掉停用词包括连词、代词、冠词及低频词等,得到候选词集(称为原始特征词集)，然后采用本文的方法对原始候选词集进行特征构造,得到结果特征词集。实验结果如表 1 所示。

表 1　特征维数约减测试结果

文档集编号	1	2	3	4	5
原始特征词数	4153	4238	4172	4451	4368
结果特征词数	2629	2615	2595	2662	2660
TP(%)	36.7	38.3	37.8	40.2	39.1

从表 1 可以看出，本文提出的文本特征描述方法可有效压缩特征空间的维数,达到比较满意的程度。

实验二：本文方法对文本分类算法性能的影响

在实验中，测试采用 5 份交叉评价的方法。针对采集到的 5 个文档集,每次选择其中 1 个文档集作为测试集，其他 4 个文档集作为训练集，这样每一个文档集轮流作为测试集，总共进行 5 次实验。以宏平均准确率 MP、宏平均召回率 MR 和宏平均 F1 值 MFl 进行对比。分类结果如表 2 所示。

表 2　本文方法对文本分类算法性能的影响

测试集编号 \ 测试值	MP		MR		MF1	
	特征生成前	特征生成后	特征生成前	特征生成后	特征生成前	特征生成后
1	0.72	0.81	0.71	0.80	0.71	0.80
2	0.77	0.87	0.75	0.86	0.76	0.86
3	0.75	0.82	0.76	0.85	0.75	0.83
4	0.78	0.86	0.75	0.83	0.76	0.84
5	0.75	0.81	0.71	0.87	0.73	0.84

从表 2 可以看出，经过特征生成后，就文本分类的 MP 值来说，2 号和 4 号测试集要高于其它测试集，分别为 0.87 和 0.86，但提高最大的是 2 号测试集，提高达 10 个百分点，提高最小的是 5 号测试集，提高了 6 个百分点；对于 MR 值来说，在不同的测试集上提高了 8 至 16 个百分点；而对于 MF1 值来说，在不同的测试集上提高了 8 至 11 个百分点。这充分说明基于词聚类的文本特征描述方法能为文本分类带来较高的精确度，其主要原因是该方法能够把语义相关的特征词聚集成一组，新的特征项不仅具有比较直观的实际意义，而且包含了原始特征空间较多的信息，使得在新特征空间上的统计信息更加可靠，较少受噪音特征的影响，可以极大的消减特征空间的维数，从而减少分类模型的训练时间与模型规模，提高分类的精度。

实验三：本文方法与传统方法的对比

传统的特征选择方法中比较著名的有文档频率(DF)、信息增益(IG)、互信息(MI)、卡方统计量(CHI)、期望交叉熵(ECE)，文本证据权(WET)，优势率(OddRatio)等方法，其中效果最好的是 IG 和 CHI[10]。

针对采集到的 5 个文档集,选择 1 号文档集作为测试集，其余作为训练集。以宏平均准确率 MP、宏平均召回率 MR 和宏平均 F1 值 MFl 进行对比。实验结果如图 1 至图 3 所示。

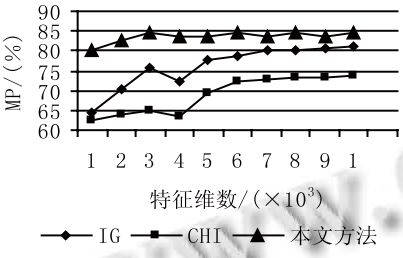

图 1 宏平均准确率 MP 比较结果

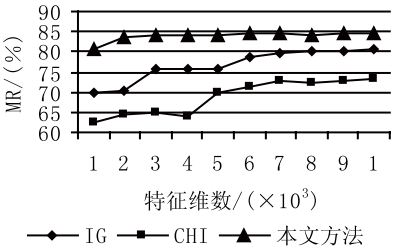

图 2 宏平均召回率 MR 比较结果

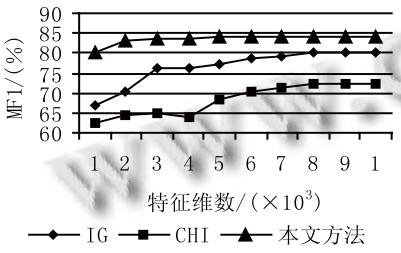

图 3 宏平均 F_1 值 MF_1 比较结果

从图 1 至图 3 可以看出，IG 方法在高维空间上的结果比在低维空间上的结果好，在 8000 维至 10000 维 MF1 能够达到约 80%，可是在 2000 维以下不超过 70%，受维度影响较大。CHI 方法随着维度的增加，分类的效果逐渐变好，5000 维以下效果较差，总体性能较差。而本文方法在向量空间从 2000 维至 lO000 维之间变化时，分类结果基本保持稳定，MP、MR 和 MFl 都能够保持在 80％至 85％之间，效果最好。其主要原因是本文方法采用了密切相关的词类表示的概念来代替具体的词，并不完全依赖词的表现形式，因而减轻了对于语料分布的依赖程度，不仅降低了特征空间的维数，而且能够获得较好的分类效果。

5 结束语

文本特征描述是文本分类的一项重要环节，直接影响到文本分类的效果。本文提出了一种基于词聚类的文本特征抽取方法，其基本思想是利用词聚类获取词汇之间的语义关联，寻找文本中的同义词，对它们进行聚类。该方法能够从语义信息角度更好地表达文本内容，化简文本表示，消除词汇分量间的同义现象，有效地降低向量维数，提高文本分类效率。结合了该方法之后的 KNN 分类算法在实验语料上的实验结果充证实了该方法的有效性，为我们今后如何更好地利用文本中的语义信息提供了新的思路。

参考文献

1 史忠植.知识发现.北京:清华大学出版社, 2002.
2 周茜,赵明生,扈雯.中文文本分类中的特征选择研究.中文信息学报,2004,18(3):17－23.
3 李蕴,罗振声,厉宇航.基于语义相关和概念相关的自动分类方法研究.计算机工程与应用,2003,39(12):106－109.
4 廖莎莎,江铭虎.中文文本分类中基于概念屏蔽层的特征提取方法.中文信息学报,2006,20(3):22－28.
5 韩客松,王永成,沈洲,吴芳芳.三个层面的中文文本主题自动提取研究.中文信息学报,2001,15(4):20－27.
6 Dhillon IS, Mallela S, Kumar R. Enhanced word clustering for hierarchical text classification. Proc. of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2002. 191－200.
7 Li H, Yamanishi K. Document classification using a finite mixture model. Proc. of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics. 1997. 39－47.
8 代六玲,黄河燕,陈肇雄.中文文本分类中特征抽取方法的比较研究.中文信息学报,2004,18(1):26－32.
9 Li H, Yamanishi K. Topic analysis using a finite mixture model. Information processing and management, 2003,39(3): 521－541.
10 Yang Y, Pedersen J. A comparative study on feature selection in text categorization. Proc. of the Fourteenth International Conference on Machine Learning. l997. 412－420.