维普资讯 http://www.cqvip.com

基于上下文关系的文本分类特征描述方法*)

何中市[1,2]　刘　里[1]

(重庆大学计算机学院[1]　重庆大学语言认知与信息处理研究所[2]　重庆 400030)

摘　要　文本特征描述是文本分类的基础，其目标是用一定的可计算的特征来表示文本，在分类的时候用这些特征来区分文本。在向量空间模型(Vector Space Model，简记为 VSM)中采用"词袋"法来处理文本，即文本被看成是由相互无关的词语构成的集合，不考虑词语之间的关系，但是这种处理方法不是很合理，因为文本的结构是完整的，孤立地对待单个词语将丢失文本的内容信息。在实际语言环境中，词语有一定的上下文"作用域"，"作用域"中的词语对表达同一主题具有一定的共性。本文提出了一种基于上下文关系的文本特征描述方法，包括特征选择方法 CBFS 及权重计算方法 CBFW。该方法是在提取一个初始特征词语集合的基础上，通过用互信息(MI)来衡量词语在上下文中的依赖度，选取对主题贡献大的词语加入特征集合，同时调整不同贡献的特征词语的权重，从而更加合理地表示文本。

关键词　特征描述，文本分类，向量空间模型，权重计算

Context Based Feature Description Model in Chinese Text Categorization

HE Zhong-Shi[1,2]　LIU Li[1]

(College of Computer[1], Institute of Langnage Recognation and Information Processing[2], Chongqing University, Chongqing 400030)

Abstract　Text feature description is considered as the basic problem in text classification and it aims to use computable feature to model documents. The most used feature description method treats a text as a set of words, which called "bag of words" model, under this model feature selection and weighting consider the "frequency" of single word only, ignoring the relation of words in context. But generally words in a certain context field can deliver correlative meaning for a same topic. So the "bag of words" model loses the context information that is important facts for improving classification precision. This paper presents a new feature description method based on text context. First, a commonly used feature selection method is used to get an initial set of feature words; secondly, Mutual Information (MI) is used to compute the word dependence in a concrete context, then, the feature words is selected according to the denpendence. Meanwhile, the weight of each feature is adjusted. Experiment result indicates the efficience of the new approach.

Keywords　Feature description, Text categorization, Vector space model, Weighting

1　引言

文本的特征描述是文本分类的一项基础性工作，它研究的是用什么样的方法和模型来表示文章的主题思想。这个描述一方面要能很好地概括文章的主要内容，另一方面要方便计算机进行计算。目前，基于矢量的方法即 VSM 得到了广泛的应用，它用若干个特征项及其权重来表示一篇文档。在这个模型中，有两个主要影响描述准确度的因素：一个是特征项的选择，一个是特征项的权重计算方式。在 VSM 中采用"词袋"方法来处理文本，即文档被看成是由相互无关的单词构成的词的集合，不考虑单词之间的依赖关系、单词出现的顺序、位置等上下文环境，着重统计每个单词在每篇文档中出现的频率，因此没能很好地表达文本的主题。本文给出了一种新的考虑上下文关系的特征描述方案，特征选择过程通过计算文本上下文"作用域"中词语的依赖度，选取对文本主题作用较大的特征词语；权重计算则改进常用的 TF-IDF 权重计算方法，突出重要特征，抑制次要特征，从而更加合理地表示了文本。

2　上下文中词语之间的关系

词语的上下文是词语在实际应用中的语言环境，它在自然语言处理中的价值体现在两个方面：①在自然语言知识获取的过程中，上下文是知识获取的来源，在相应推理机制下，上下文本身就是知识；②在自然语言处理的应用问题解决过程中，上下文扮演着解决问题所需信息和资源提供者的重要角色。因此，当前文本分类中，采用"词袋"方式来考虑每一个单独的词语，本来就丢失了词语间的这种关系，因此有必要把单个特征词语和它"作用域"中的词语同时加以考虑。

通常情况下，上下文的选取是基于核心词左右一定范围进行的，这个固定的范围被称为"窗口"，表示为[a,b]，即核心词语左 a 个位置和右 b 个位置的范围。文[1]分别对中文和英文文本核心词语的作用域进行了研究，得出的上下文信息如图 1 所示。

下面将给出上下文的定量描述：

上下文位置 p 的信息量 IG_p，是整个系统熵 $H(W)$ 相对于已知上下文位置 p 时整个系统条件熵 $H(W|V_p)$ 的减少量，即信息增益。信息增益如下计算：

$$IG_p = H(W) - H(W|V_p)$$

$H(W)$ 是上下文中的核心词，w 为信源的信息熵：

$$H(W) = -\sum_{w \in W} p(w) \times \log_2 p(w)$$

其中 $P(w)$ 为核心词 w 的词频(统计频率)，定义为

$$P(w) = \frac{fre(w)}{n}$$

$H(W|V_p)$ 为已知上下文位置 V_p 的条件熵，定义为

$$H(W|V_p) = \sum_{cw \in V_p} p(cw) \times H(W|cw)$$

其中 $P(cw)$ 为上下文位置 p 中的上下文词语 cw 的词频；H

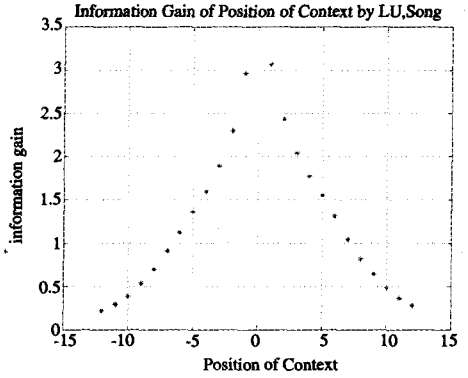
*)国家自然科学基金项目(60173060)。**何中市**　教授，博导，主要从事自然语言处理、机器学习与数据挖掘的研究；**刘　里**　硕士研究生，主要研究方向为机器学习、自然语言处理。

维普资讯 http://www.cqvip.com

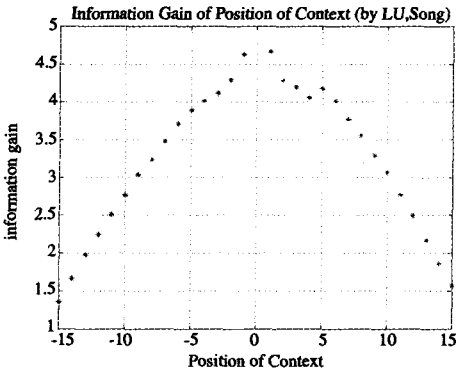
$(W|cw)$是在上下文词语已知情况下的条件熵：

$$H(W|cw)=-\sum_{w\in W}p(w|cw)\times\log_2 p(w|cw)$$

由此可得上下文各位置对核心词的贡献大小，即信息量。文[11]通过对《人民日报》统计汉语词语最近距离左 8 个位置和右 9 个位置的上下文范围，可为确定词语提供 85%的信息量。



(a) 中文上下文位置与其信息量的关系



(b) 英文上下文位置与其信息量的关系

图 1　上下文位置和信息量的关系

3　基于上下文的特征选择方法(CBFS)

常见的特征选择方法有：文档频率(DF)、信息增益(IG)、互信息(MI)、χ^2统计量等[2]。这些方法的基本思想都是对每一个特征即词条，计算它的某种统计的度量值，然后设定一个阈值 T，把度量值小于 T 的那些特征过滤掉，剩下的即认为是有效特征。这些特征选择方法都是孤立地考虑每一个特征词语对主题的贡献，没有考虑到特征词语出现的具体语言环境。特别是在中文文本中，同一个词在不同的语言环境中表现出不同的语义[3]，所以用这些特征进行文本分类没能体现真实的语言环境。本文提出了一种基于上下文的特征选择方法 CBFS，通过计算文本特征词语间的依赖关系，选取相关度大的特征，从而体现上下文中词之间的语义相关性，能较准确地描述文本的主题。

目前，文本的特征描述大都采用向量空间模型(VSM)，把每篇文档映射成为向量空间中的一个向量，就可以用数学的方法对文本进行处理。VSM 模型没有考虑每个词出现的具体语言环境和词的顺序。于是有学者考虑用图或树的模型来描述文本主题，考虑了特征词的出现顺序。但是在计算文本间的相似性时却存在一个问题：什么样的两棵树或者两个有序图才是相似的。如果用这种表示方法，计算开销又是否可以满足要求。这些都是 VSM 当前不可替代的原因。综合上述考虑，本文在文本表示模型上仍然采用 VSM，但力求在该模型上对文本的上下文信息进行一定的量化。

本文提出的基于上下文关系的特征选择方法，是在应用一般的特征选择方法(例如 文档频率 DF、信息增益 IG、互信息 MI)得到一个初始特征集合之后，把这些特征返回到原文，通过计算上下文中特征词语间的依赖度，对这些特征词语进行进一步的处理，选取依赖度大的特征词语，并加大这些词语在分类时的作用，从而在一定程度上考虑特征之间的关系，加入特征词语在上下文中依赖关系的量的描述。图 2 描述了考虑上下文的特征选择方法。

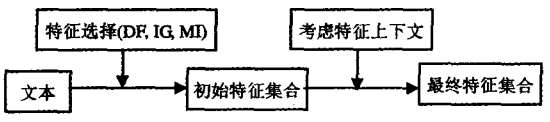


图 2　考虑上下文的特征选择方法

在考虑上下文的作用之前，先对初始特征集合中的特征词语做个区分，把出现在文本的标题、摘要中的特征称为“关键词语”。文中把关键词语标记为 k，关键词语的集合记为 K。在考虑上下文信息的时候，对这部分特征进行特殊处理。同时对文本的各个部分的重要程度(在表达文本内容时的贡献大小)做个区分：无论哪种类型的文本，一般都会有标题，通常标题部分特征信息的量都比文章其余部分要大；摘要部分含有的信息量次之；文章的开头和结尾含有的信息量比正文的其它部分要大[4~7]。根据本文后面章节的内容，对文本的各个部分进行了如下表示：文本的标题(T)、副标题(S)、摘要(A)及文本开头和结尾(F)。为了体现特征词语作用的差别，给特征词语赋予“重要系数”，权重计算时用于区分它们对文本内容的贡献大小的差别。对 T，S，A 中的关键词语分别赋予重要系数 IT，IS，IA，对 F 区域和正文中其它位置的特征词语赋予重要系数 IF，ID，且 IT>IS>IA>IF>ID，充分体现特征词语在文本的不同区域对主题内容贡献的大小。

本文在计算上下文中特征词语的依赖度时分为如下两步：

1)计算“关键词语”的上下文中的词语对“关键词语”的依赖程度，提取那些对“关键词语”依赖度较大的词语，加入特征集合中。

2)计算上下文中特征词语之间依赖关系，提取相互依赖较大的特征词语。

词语之间的依赖程度如何定量地描述呢？本文利用互信息(Mutual Information)来描述词语在上下文中的关系，词语间的互信息量衡量了词语之间依赖度。词语 x，y 的互信息计算公式如下：

$$I(x,y)=\log\frac{P(x|y)}{P(y)}$$

考察对象：上下文中特征词语。

评估方式：互信息量。

根据互信息的公式，得到词语 w_i 和特征词语 k 之间的互信息计算公式：

$$I(w_i,k)=\log_2\frac{P(w_i|k)}{P(k)}$$

其中，w_i 表示在关键词语 k 作用域$[a,b]$内的词语。$P(w_i|k)$ 表示特征词语 k 作用域内词语 w_i 出现的概率。特征词语的词频统计概率 $P(k)$计算如下：

$$P(k)=\frac{fre(k)}{n}$$

维普资讯 http://www.cqvip.com

最终，选取互信息量大于零的词语加入到文本的特征集合中。通过考虑上下文的信息而选取的特征词语，记为 K'。

4 考虑上下文的特征权重计算方案(CBFW)

通过前述特征选择之后，得到了一个最终的特征词语集合，其中包含关键词语集合 K，以及在上下文中选取的特征词语集合 K'，它们在呈现文本内容时起核心作用。在提取这些词时，没有采用传统的基于单个词语频率的提取方法，而是考虑到这些词语在文本中特殊的结构特征以及词语作用域内的互信息量。运用前面的策略，假设得到一个文本的特征向量为

$$Term=\{t_1,t_2,t_3,\cdots,t_n\}$$

对于特征 $t_i\notin(K\cup K')$，用 *TF-IDF* 函数来计算权重，得到权重的最大值为

$$G=\underset{t_i,d}{\mathrm{Max}}(TFIDF(t_i,d))$$

其中特征权重 $TFIDF(t,d)$ 如下：

$$TFIDF(t,d)=\frac{tf(t,d)\times\log_2(N/n_t+a)}{\sqrt{\sum_{t\in d}[tf(t,d)\times\log_2(N/n_t+a)]^2}}$$

为了区分特征词语对文本主题的不同贡献，本文提出如下特征的权重公式：

$$Score(t_i,d)=\frac{\alpha\times G\times tf(t_i,d)\times\log_2(N/n_t+0.01)}{\sqrt{\sum_i[\alpha\times G\times tf(t_i,d)\times\log_2(N/n_i+0.01)]^2}}$$

其中，参数 α 为前文说明的“重要系数”。当 $t_i\notin(K\cup K')$ 时，$\alpha=I_D=1/G$；当 $t_i\in(K\cup K')$ 时，根据关系 $I_T>I_S>I_A>I_F>I_D$，赋予 α 不同的值，目的是为了区分文本的不同部分的重要程度，系数的具体大小可以通过实验来给定。对于存在于不同区域的同一特征，则取其所有权值中的最大值。

从上述权重计算方案可以看出，我们实现了对 VSM 模型的特征词权重的改进，对贡献大的词的权进行了加强。新得到的文本特征描述不但具备词在文本中的频率信息，还蕴涵了词的上下文环境，突出了该文本的语义环境信息，反映了真实的语言特性。

5 实验过程及结果分析

5.1 实验数据

实验数据来源于复旦大学国际数据库中心自然语言处理小组，训练语料是由人工标注的 9804 篇文档，分为 20 个类别，其中包括计算机、体育、艺术、教育、能源等；测试语料共 9833 篇文档，训练语料和测试语料基本按照 1:1 的比例来划分。

5.2 LIBSVM 分类器

LIBSVM 是台湾大学林智仁(Lin Chin-Jen)副教授等开发设计的一个简单、易于使用和快速有效的 SVM 模式识别与回归的软件包，软件有一个特点，就是对 SVM 所涉及的参数调节相对比较少，提供了很多的默认参数，利用这些默认参数就可以解决很多问题。该软件包可以在 http://www.csie.ntu.edu.tw/~cjlin/获得。LIBSVM 使用的训练数据和测试数据文件格式如下：

〈label〉〈index1〉:〈value1〉〈index2〉:〈value2〉…〈indexn〉:〈valuen〉

其中〈label〉是训练数据集的目标值，对于分类，它是标识某类的整数(支持多个类)；对于回归，是任意实数。〈index〉是以 1 开始的整数，可以是不连续的；〈value〉为实数，也就是我们常说的自变量。测试数据文件中的 label 只用于计算准确度或误差，如果它是未知的，只需用一个数填写这一栏，也可以空着不填。

5.3 实验方案

本文提出的文本特征描述方案涉及到特征选择与权重计算两个方面，所以在实验中，分别对特征选择与权重计算进行了独立的实验。其中特征选择实验对比了信息增益和本文提出的 CBFS 两种特征选择方法。权重计算则比较了 TF-IDF 法和本文提出的 CBFW 权重计算方法。实验中均采用 LIBSVM 作为分类器。最终的实验结果通过查全率(recall)和查准率(precision)两个指标加以衡量，R 代表查全率，P 代表查准率。

下面通过一个训练文本来说明整个实验过程。选择类别为“计算机类”的一篇标题为“一种基于智能 Agent 的协同工作模型”的训练文本，先通过分词，去掉停用词等预处理。先采用信息增益(IG)的特征选择方法得到一个初始特征词语集合(表 1 中非斜体的词语)，然后采用基于上下文关系的特征选择之后得到的文本的最终特征词集合(表 1 中包含了斜体的词语)，用 TF-IDF 法和 CBFW 法分别计算得到特征词语的权重，得到表 1 所示的特征向量。表中“标记”栏中“KeyWord”代表在文本标题和摘要处得到的特征词语，“Context”代表考虑了上下文后得到的特征词语，“*”表示考虑上下文后无变化的词语。

5.4 实验结果及分析

表 2 不同特征选择方法分类结果比较

	特征选择方法	通讯	环境	农业	经济	法律
R	IG	.788	.817	.802	.742	.783
	IG+CBFS	.817	.852	.880	.835	.812
P	IG	.801	.823	.779	.769	.801
	IG+CBFS	.826	.842	.868	.830	.822

表 3 不同权重计算方法分类结果比较

	权重计算方法	艺术	文学	教育	哲学	能源
R	TF-IDF	.832	.864	.885	.845	.789
	CBFW	.904	.877	.910	.884	.847
P	TF-IDF	.832	.830	.803	.763	.752
	CBFW	.906	.901	.873	.821	.796

本文进行的对比实验，得到的部分实验结果如表 2 所示。在对本文提出的特征选择方法 CBFS 和 IG 进行比较时，同时采用 TF-IDF 这种权重计算方法；而在对本文提出的权重计算方法 CBFW 和 TF-IDF 进行比较时，都采用本文提出的特征选择方法 CBFS。实验中，通过对实验数据的分析，最终把重要系数设置为 $I_T:I_S:I_A:I_F:I_D=1.0:1.0:0.8:0.6:0.3$。部分实验结果如表 2 和表 3 所示。

通过对试验结果进行分析发现，CBFS 方法和 CBFW 方法在查全率和查准率上较其它方法都有所提高，事实上：

1)新的特征选择方法 CBFS 通过加入了上下的信息，选取了和主题高度相关的特征，所以文本向量表达主题更加贴切。

2)新的权重计算方法 CBFW 通过对重要程度不同的特征的权值进行了调整，使得重要特征得以突出，次要特征得到抑制，更能体现文本的主题。

结束语 文本特征描述是文本分类的一项重要环节，直接影响到分类的效果。现有方法由于把文本当成“词袋”处理，即文档被看成是由相互无关的词语构成的集合，不考虑词语之间的上下文关系、词语出现的顺序等，着重统计每个词语在每篇文档中出现的频率，因此没能很好地表达文本的主题，

维普资讯 http://www.cqvip.com

表1 文本特征样例

特征	标记	特征频数	文档频数	TF-IDF	CBFW
智能	KeyWord	1	2	0.00481840732768579	0.0061249452298419
协同	KeyWord	1	2	0.00481840732768579	0.0061249452298419
模型	KeyWord	1	66	0.00143062061709339	0.00181854134125023
计算机	KeyWord	6	3	0.0265520035058774	0.0337517266922674
研究	KeyWord	9	236	0.00180874631054646	0.00229919791610821
开放	*	1	2	0.00481840732768579	0.0061249452298419
扩充	***Context***	***2***	***28***	***0.00452131952232129***	***0.0057473004164082***
结构	Context	25	178	0.0118121259020116	0.0150150494297389
分布式	Context	2	2	0.00963681465537158	0.0122498904596838
人工	*	5	27	0.0114794266805007	0.0145921369669304
系统	***Context***	***9***	***165***	***0.00491018965418567***	***0.00624161484381356***
技术	Context	5	170	0.00258400445043399	0.00328467160541559
特征	***Context***	***8***	***126***	***0.00644571931422172***	***0.0081935118975306***
设计	*	6	121	0.00506882731103197	0.0064432679821838
应用	*	4	133	0.00301412073069486	0.00383141630338334
涉及	*	1	62	0.00149110107960332	0.000189542141700064
群体	Context	23	34	0.047670331993738	0.0605964072137211
分布	Context	3	60	0.0045684729800506	0.00580723979603064
控制	Context	16	106	0.015561906085527	0.0197816033314841
集合	*	5	41	0.00945676443959949	0.00120210186281372
消息	Context	2	33	0.00420306517977912	0.0053427496417093
被动	*	1	25	0.00237043343995017	0.00301318962953962
主动	*	1	46	0.00177995878404261	0.00022626044919438
交互	Context	5	38	0.00392984398412891	0.00499544300173845
检查	*	1	39	0.00193977182292144	0.00024657517237113
通信	*	2	2	0.00963681465537158	0.00122498904596838
数据	Context	3	122	0.00251057528963335	0.00319133171993247
同步	*	1	39	0.00193977182292144	0.00024657517237113
算法	Context	2	41	0.00378270577583979	0.00548084074512548
类型	*	3	81	0.00369771521790728	0.0027003712316104
规则	*	3	61	0.00452049722769163	0.00374625515202373
适应性	Context	2	13	0.00400402539773197	0.00481858356483103
效率	*	1	50	0.00169925296729905	0.00021600148448536
状态	*	26	130	0.0201642731326202	0.02563192775904
通讯	Context	36	13	0.108144914318351	0.137469008333917
层次	Context	7	99	0.00727034897970527	0.00924174447574549
分解	Context	2	27	0.00459177067220026	0.00583685478677218

不能很好地概括文本蕴涵的语义信息。本文利用组成文本的词语之间存在潜在的联系，即词语有一定的上下文“作用域”，“作用域”中的词语对表达同一主题存在一定的共性，提出了基于上下文关系的特征描述方法，包括特征选择以及权重计算，考虑了上下文中词语之间的关系，融入了文本的内容信息，体现了真实的语言特性，通过实验验证了该方法的性能。

参 考 文 献

1 鲁松，白硕. 自然语言处理中词语上下文有效范围的定量描述[J]. 计算机学报，2001，24(6)：742～747

2 任纪生，王作英. 基于特征有序对量化表示的文本分类方法. 清华大学学报(自然科学版)，2006，46 (4)：527～529

3 刘开瑛，薛翠芳，郑家恒，等. 中文文本中抽取特征信息的区域与技术. 中文信息学报，1998，12(2)：1～7

4 Angheluta R，De Busser R，Moens M-F. The use of topic segmentation for automatic summarization. In：Hahn U，Harman D，eds. Proceedings of the Workshop on Automatic Summarization. Philadelphia，Pennsylvania，USA，2002. 66～70

5 谌志群，张国煊. 文本挖掘研究进展. 模式识别与人工智能，2005，18(1)：65～74

6 Ko Y，Park J，Seo J. Improving text categorization using the importance of sentences. Information Processing and Management，2004，40：65～79

7 Sebastiani F. Machine Learning in Automated Text Categorization. ACM Computing Surveys，2002，34(1)：1～47