基于领域词语本体的短文本分类

宁亚辉 樊兴华 吴 渝
(重庆邮电大学计算机科学与技术研究所 重庆 400065)

摘 要 短文本自身长度较短,描述概念能力弱,常用文本分类方法都不太适用于短文本分类。提出了基于领域词语本体的短文本分类方法。首先抽取领域高频词作为特征词,借助知网从语义方面将特征词扩展为概念和义元,通过计算不同概念所包含相同义元的信息量来衡量词的相似度,从而进行分类。对比实验表明,该方法在一定程度上弥补了短文本特征不足的缺点,且提高了准确率和召回率。

关键词 短文本,本体,知网,文本分类,语义,义元
中图法分类号 TP18 **文献标识码** A

Short Text Classification Based on Domain Word Ontology

NING Ya-hui FAN Xing-hua WU Yu
(Institute of Computer Science and Technology, Chongqing University of Postsand Telecommunications, Chongqing 400065, China)

Abstract The conventional methods of text classification are not suitable for short text classification because short texts are short and their ability of describing concept is weak. A method using the domain word ontologies for short texts classification was proposed. First, the domain high frequency words were got as the feature words. Then the feature words were extended to concept and "sememe" by Hownet, which extended the feature from semantic and amended the feature scarcity. Last, the word similarity values were got by calculating the same "sememe" message included between different concepts. Experimental results prove that the classification efficiency and precision are both improved.

Keywords Short-text, Ontology, Hownet, Text-classification, Semantic, Sememe

1 引言

短文本分类就是对内容较短的文本(通常少于160个中文字符)进行分类。国内外对于文本分类[1,2]进行了多年的研究,但在短文本分类领域,目前国内外的研究工作较少[3,4],且效果不是很理想。Zelikovitz, S 等学者提出了将LSI应用于短文本分类的特征选择[3]。LSI(潜在语义索引)是一种常用的特征抽取算法,其核心技术是矩阵理论中的奇异值分解。LSI在处理文本数据高维稀疏问题上比较成功,但是不能很好地解决多义词的问题。

传统的文本分类方法也不能很好地应用到短文本分类中。例如:利用互信息[1](mutual information, MI)进行特征选择,是对词和类别相关性的度量,而短文本中词的个数少而且信息弱,有必要考虑其它语义关系(如:同义关系或近义关系等)。传统的贝叶斯分类器(Naïve Bayes)在知识储备不足的情况下容易对新增未标注的训练文本产生分类错误,若这些分类错误的文本过早地加入到分类器中会降低其分类性能[5]。本质原因在于短文本长度较短,描述概念的信号弱,词的个数少等缺点导致短文本特征严重不足,使得传统的分类器分类效果也不好。

由于本体强调相关领域的本质概念,同时也强调这些本质概念之间的关联,而短文本具有特征不足的缺点,有必要强调概念之间的语义关系,这和本体表示基本一致。因此,本文提出了基于领域词语本体的短文本分类方法。该方法以在领域出现的比重大于阈值的领域高频词作为领域之间互相区别的特征词,但是如果阈值过大,特征词的领域区别性不好,如果阈值过小,特征词数量有限会使得分类能力降低,因此,为了保证特征词良好的领域区别性并且使得分类能力提高,本文借助知网将特征词扩展为概念,概念扩展为义元,从而得到扩展后的领域词语本体,有效地从语义方面对特征进行扩展,使得原来的特证词不再是孤立的词,而是由概念和义元支撑的相关词的集合,一定程度上弥补短文本特征不足的缺点;最后通过计算不同概念所包含相同的义元信息量来衡量概念之间的相似程度,词之间的相似度又可以通过概念间相似度获得,并以此作为待分类短文本和领域词语本体相似度比较的基础从而达到分类的目的。

本文在第2部分介绍了传统的文本分类过程和本文提出的短文本分类过程;第3部分给出了领域词语本体的初步框架和获取;第4部分给出了相似度计算方法;第5部分进行了实验并做了结果分析;最后提出了进一步的工作。

到稿日期:2008-06-13 本文受国家自然科学基金(60703010),重庆市自然科学基金(2006BB2374),重庆市教委科学技术研究项目(KJ070519),教育部回国留学人员启动基金(教外司留[2007]1109号)资助。
宁亚辉(1980—),女,硕士研究生,主要研究方向为自然语言处理、数据挖掘,E-mail:ningyh@126.com;樊兴华(1972—),男,博士,教授,主要研究领域包括人工智能、自然语言处理、信息检索;吴 渝(1970—),女,博士,教授,主要研究领域包括数据挖掘、网络智能和机器视觉。

2 短文本分类的实现过程

传统的文本分类流程见图 1。其特征选择就是选取适合分类的特征词，目前的方法较多[1]。典型的利用互信息进行短文本特征选择，如式 (1)所示：

$$MI(t)=\sum_{i=1}^{|C|}P(c_i)\log\frac{P(t|c_i)}{P(t)} \tag{1}$$

$|C|$表示文本的类型数，$P(c_i)$表示第 i 类文本在训练文本集合中出现的概率，$P(t)$表示词 t 在训练文本集合中出现的概率，$P(t|c_i)$表示在第 i 类文本中 t 出现的概率。MI 越大，词和类的共现程度越大。

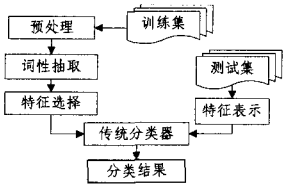

图 1 传统文本分类过程

传统的文本分类器参考文献[1]，典型的贝叶斯分类器(Naïve Bayes)如式(2)：给定二值文本向量 $d=(W_1,W_2,\cdots,W_D)$，$W_i=0$ 或者 1。如果第 i 个特征出现在文本中，$W_i=1$，否则 $W_i=0$。令 $P_{ki}=P\{W_k|c_i\}$，$P\{\cdot\}$表示$\{\cdot\}$发生的概率。

$$P(c_i|d_j)=\log P\{c_i\}+\sum_{i=1}^{D}\log(1-P_{ki})+\sum_{i=1}^{D}W_k\log\frac{P_{ki}}{1-P_{ki}} \tag{2}$$

本文的短文本分类流程见图 2。这种方法有以下两个方面的特点：

(1)本文的领域词语本体是从不同级别给出了词的表述。这样就可以进行不同级别的语义相似性比较，较多地考虑了词与词之间的语义关系，并且通过借助外部资源可以更新所建立的领域词语本体，从相似度比较来说，也有利于更深层次的考虑和扩展语义关系。改变了传统的单一的关键词匹配方式，这种方式对短文本较为适用，尤其是对近义词和同义词以及同一话题下的相关词这种情况比较有效。

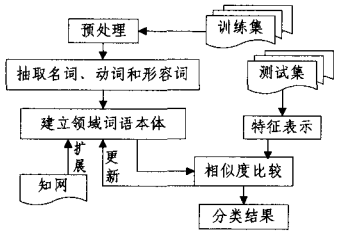

图 2 本文短文本分类过程

(2)本文所建立的词语本体有很高的可重用性和可扩充性。可重用是因为本体中的特征词都是在某个领域出现而在其他领域出现几率很小的词的集合，在任何情况下，这些本体都是可用的。可扩展是因为通过相似性比较，我们可以把相似度大于某个阈值的词添加到领域词语本体中进行扩充。这样，领域词语本体就具有可重用和可更新的优势。

3 领域词语本体

3.1 本体(Ontology)

本体在哲学界和人工智能领域都有其不同的定义。在人工智能领域，1998 年 Studer 给出了共享概念模型的明确的形式化规范说明[6]。

本体的目标是捕获相关领域的知识，提供对该领域知识的共同理解，确定该领域内共同认可的词汇，并从不同层次的形式化模式上给出这些词汇(术语)和词汇之间相互关系的明确定义。某个领域的本体就是关于该领域的一个公认的概念集，其中的概念含有公认的语义，这些语义通过概念之间的各种关联体现，本体通过它的概念集及其所处的上下文刻画概念的内涵。总之，本体强调相关领域的本质概念，同时也强调这些本质概念之间的关联。考虑到短文本自身的缺点，有必要强调短文本分类时概念之间的语义关系，这和本体表示基本一致，这为本文提出的基于知网扩展的领域词语本体的短文本分类方法提供了理论依据。

知网是一个以汉语和英语的词语所代表的概念为描述对象，以揭示概念与概念之间以及概念所具有的属性之间的关系为基本内容的常识知识库。知网也可以称之为一种本体表示，对于知网的详细介绍可见文献[7]。

3.2 领域词语本体表示

定义 1(领域词语本体) 领域词语本体是指某个领域最核心最本质的概念集合并且给出这些概念正规明确的表述。这些概念在其他领域出现的几率很小并且可以唯一确定某个领域。

在具体的应用中，本体的表示方式可以多种多样[8]。可以用自然语言、框架、语义网络或逻辑语言等来描述本体，目前比较形式化的本体表示有 Ontolingua，CycL 和 Loom 等，都是基于逻辑语言的本体模型。但是目前尚没有一个标准的本体构造方法，在已有的本体模型中，出于对各自问题域和具体工程的考虑不同，构造本体的过程和方法也各不相同。尽管逻辑语言的表达能力很强，但其推理过程比较复杂，不利于本体模型的实施。因此针对短文本分类的具体情况，本文结合知网提出了一种类框架结构的领域词语本体描述形式，初步设计的领域词语本体框架如图 3 所示。

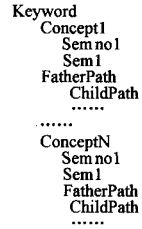

图 3 领域词语本体框架示意图

3.3 领域词语本体的获取

领域词语本体的获取有两个部分：第一，获得领域高频词，所谓领域高频词是指在某个领域出现几率大而在其他领域出现几率小的词(即就是领域区分能力强的词)，这些领域高频词可以唯一确定某个领域，所以可以用这些高频词作为特征词来进行分类，但是其数量有限并且没有体现语义关系，因此难以解决短文本中特征不足的缺陷。第二，利用知网对高频词进行概念和义元级别的扩展，使得原来的特证词不再是孤立的词，而是由概念和义元支撑的相关词的集合。在某个待比较的词与特征词不同的情况下，可以进行义元级别的比较，从而可以更精确地确定两者之间的相似性。领域词语本体的获取具体步骤如下：

Step1　对训练语料进行预处理：分词及词性标注；通过词性选择，过滤掉对分类作用不大的词，仅保留名词、动词和形容词。

Step2　进行词频统计，得到每个词在不同类中出现的次数。过滤掉所有类中出现次数均不超过 3 的词，对每个词在不同类的比重进行归一化处理。假设“*keyword* $x_1x_2\cdots x_u$”表示 *keyword* 这个词在领域 i 中出现的次数为 x_i，归一化后得到 *keyword* 在领域 i 中的比重为 $x_i/\sum_{j=1}^{u}x_j$。

Step3　抽取在每类中比重不小于 *weight* 的词作为领域高频词。在此基础上，人工过滤掉一些和本类关联不大的词，得到了初始的未经过知网扩展的领域本体词库。

Step4　借用知网来扩展初始领域本体词库，本体词库中的每个 *keyword* 在知网找到对应的所有概念（义项）得到：*keyword*（概念 1| 概念 2| …|概念 n），*keyword* 为特征词，由不同的概念组成，每个概念又由不同的义元组成：格式是：“语义号 义元 1 …义元 m ”。由于很多概念只是语义号不同，而包含的义元是完全相同的，为了简化后面的相似度计算的复杂度，过滤掉重复的概念。

Step5　将概念中的义元扩展为：“义元 义元号 父路径|子路径”。“义元号”表示“义元”在整个知网树中的节点深度；“父路径”给出了“义元”的上层义元；“子路径”给出了“义元”的下层义元，它更精确地体现了义元在知网中位置，也为词进行语义相似度比较提供了基础。

按照上述步骤，得到经过知网扩展后的领域词语本体。

我们举例说明上述过程，例如：初始本体词库中有“经济”这个特征词，经过知网扩展，由四个概念组成：“经济 046979 属性值 举止 俭 良|046980 属性值 举止 俭 良|046981 事务 工 农 商|046982 事务 工 农 商”。分别考虑每个概念对于大量的短文本分类没有多大的实际意义并且会增加大量的计算量，因此过滤掉重复概念后就由两个概念组成：“经济 属性值 举止 俭 良| 事务 工 农 商|”。每个概念由不同的义元组成，如：“属性值”、“举止”等。再对概念中的每个义元进行扩展，例如：义元“属性”被扩展为：“属性值 14 Hownet|appearance, measurement, property, relationship, situation”，其中，“14”为“属性值”对应的义元号，“Hownet”为父路径，“appearance, measurement, property, relationship, situation”为子路径。

4　相似度计算

语义相似度计算是目前研究的热点，在基于知网的相似度计算中，刘群、李素建[9]依据义元的层次体系（上下位关系），通过计算节点路径获得义元相似度。李峰，李芳[10]在此基础上又引入了节点在树中的深度，层次越深的节点的相似度越大。本文也是将词扩展为概念，概念又扩展为义元，但是在具体相似度计算时并没有采用刘群、李素建[9]的方法，原因在于刘群、李素建[9]考虑了义元的很多细节，而本文是针对短文本分类的，最关心的是词和类之间的相关性，比如“医生”和“患者”按照刘群、李素建[9]的方法相似度很低，而按照分类的目的来说，两个词应该具有很高的相似度值。因此，本文只是通过计算不同概念所包含的相同的义元信息量来衡量词之间的相似程度，当不同的概念包含的相同义元越多，概念相似度值就越大。以此作为待分类短文本和领域词语本体相似度比较的基础。这样我们就可以省略掉很多对分类意义不大的大量计算过程。例如：“战斗力 属性 力量 争斗 军队 军|”和“制海权 属性 势力 控制 水域 国家 军队|”中，“战斗力”和“制海权”是两个不同的特征词，这两个特征词各有一个概念而且是不同的，但是两个概念中有相同的义元“属性”和“军队”，从而可以确定相似度。

为了叙述方便，对文中的一些符号进行约定：待分类的短文本为 D，领域词语本体为 O_i，$\mathrm{sim}(A,B)$ 表示 A 和 B 的相似度值。具体算法如下：

Step1　词相似度比较的原则。如：D 中的词“w_k 概念 1|…概念 n|”和 O_i 中的词“kw_j 概念 1|…概念 m|”在进行相似度比较时，遵循以下规则：

Rule1　当 w_k 和 kw_j 相同时，则 $\mathrm{sim}(w_k,kw_j)=1$，并且结束 w_k 和 O_i 中其余词的比较。若不相同时，进行 Rule2；

Rule2　当 w_k 和 kw_j 不相同时，进行基于词的概念的比较，即 w_k 和 kw_j 中的概念两两进行比较，若存在两个概念完全相同，则 $\mathrm{sim}(w_k,kw_j)=\frac{f\times\alpha}{m\times n}$，否则，进行 Rule3。其中，$f$ 为完全匹配的概念数，α 为人为给出的权重因子，m 和 n 分别为 w_k 和 kw_j 的概念个数。

Rule3　当概念匹配不成功时，进行概念中基于义元的匹配比较，则 $\mathrm{sim}(w_k,kw_j)=\frac{f\times\beta}{x\times y}$。其中，$f$ 为完全匹配的义元数，β 为人为给出的权重因子，x 和 y 分别为 w_k 和 kw_j 的义元个数。

Rule4　当 $\mathrm{sim}(w_k,kw_j)$ 的值大于某个给定的阈值时，就将词 w_k 添加到词语本体 O_i 中，对于词语本体进行更新。

Step2　词 w_k 与本体 O_i 的相似度值取 w_k 和 kw_j 相似度的最大值见式(3)：

$$\mathrm{sim}(w_k,O_i)=\max\{\mathrm{sim}(w_k,kw_1),\mathrm{sim}(w_k,kw_2),\cdots,\mathrm{sim}(w_k,kw_M)\} \tag{3}$$

M 为领域本体中特征词的个数。

Step3　短文本 D 与本体 O_i 的相似度结果如式(4)：其中，N 为短文本 D 中的词的个数。

$$\mathrm{sim}(D,O_i)=\sum_{k=1}^{N}\mathrm{sim}(w_k,O_i) \tag{4}$$

最后，短文本 D 属于相似度结果 $\mathrm{sim}(w_k,O_i)$ 最大的那个类别。

5　实验及结果分析

5.1　性能评估指标

对实验分类结果的性能我们采用如下 3 个常规性能指标进行评估：

精确率(Precision,P)

$$\mathrm{P}=\frac{\text{正确分为某类的文本数}}{\text{测试集中分为该类的文本总数}}\times 100\% \tag{5}$$

召回率(Recall,R)

$$\mathrm{R}=\frac{\text{正确分为某类的文本数}}{\text{测试集中属于该类的文本总数}}\times 100\% \tag{6}$$

F-measure (F1)

$$\mathrm{F1}=\frac{2\times\mathrm{P}\times\mathrm{R}}{\mathrm{P}+\mathrm{R}} \tag{7}$$

5.2　实验评估及分析

本文以在搜狐、新浪等各大网站的不同类别的跟帖中的评论为语料集。本次实验选取其中的 82034 个短文本语料，

其中女性话题有17160篇、军事话题有16283篇、房地产话题有14878篇、国内新闻为16419篇、体育类话题为17290篇，为实验语料。

首先随机选取女性类4585篇、军事类4343篇、房地产类4350篇、国内是4259篇和体育类4064篇为实验训练集语料，其余为测试集。对训练集执行3.1节中的操作，得到经过知网扩展后的领域本体词库。对测试集短文本进行预处理，抽取名词、动词和形容词。在上述语料基础上，本文进行了三组实验。实验结果比较如表1所列。

表1　三组实验结果对比

	精确率 P(%)			召回率 R(%)			F1值(%)		
	实验1	实验2	实验3	实验1	实验2	实验3	实验1	实验2	实验3
女性	53.38	52.21	78.63	74.49	56.96	54.20	62.19	54.48	64.16
军事	62.48	43.77	52.90	43.44	40.47	55.73	51.25	42.05	54.28
房地产	52.43	57.18	68.53	85.40	66.05	80.95	64.97	61.30	74.22
体育	86.93	57.39	78.63	39.78	52.16	54.20	54.58	54.65	64.18
国内新闻	41.42	30.76	53.33	36.93	30.76	36.07	39.04	32.04	43.03
平均值	59.33	49.28	66.40	56.01	49.28	56.23	54.41	48.90	59.97

实验一(传统的基于互信息的特征选择和贝叶斯分类)

利用传统的互信息式(1)进行特征选择和多类朴素贝叶斯分类器式(2)进行短文本分类，我们进行了4组测试，分别选取经过互信息排序后的前1500，2000，2500和3000个特征词进行实验，其中效果最好的是选2500个特征。

实验二(基于未扩展的领域词语本体的短文本分类)

将每个短文本和未经过知网扩展的初始领域词语本体(3.3节中step1－step5)进行词匹配，统计短文本中出现在每类初始本体中的词的个数，短文本属于统计的词个数最多的那个类。给定 $weight=0.6$，所得结果不是很理想，我们还选择了 $weight$ 值为0.4和0.5的结果，虽然比0.6有所提高，但是分类能力均不能超过实验一，只有 $weight=0.4$ 时候的结果接近实验一。原因在于利用未经扩展的词语本体过滤掉很多词汇，只保留了每个类别中的高频词汇进行相似度匹配，并且没有考虑语义关系。因此，我们考虑用实验三进行改进。

实验三(基于知网扩展后的领域词语本体的短文本分类)

本实验是在实验二的基础上，以实验二 $weight=0.6$ 时为基准所作的改进。首先，对实验二中的测试集短文本再执行3.3节中的step4和step5的操作，然后按照第4节的算法进行相似度结果计算。实验中，人为确定 $\alpha=0.4$，$\beta=0.2$(主要是让特征词的匹配具有足够的优越性，所以概念匹配因子 α 和义元匹配因子 β 的值取得比较小)。

通过实验三的结果可以看出，经过扩展后的领域词语本体应用于短文本分类结果较实验二和实验一的准确率、召回率和F1值都有所提高。实验三所用的方法从语义方面考虑了词之间的关系，而实验一和实验二只是词匹配概率的一般统计。比如："兵 军队|人 争斗 军|"是军事类短文本中的一个词，在未扩展的军事领域本体词库中没有"兵"这个词，因此认为它和军事类的相似性为零。而在经过知网扩展后的领域本体词库中有"军人 人 争斗 军|"，两个特征词中存在完全相同的概念，因此就可以计算"兵"和"军人"的相似度从而确定和军事类的相似度。经过对实验二错误分类的短文本观察后发现这种现象比较普遍，因此，考虑了语义关系后，分类能力比实验二有较大的提高，相对传统的实验一的结果也有一定提高。

结束语　本文提出了基于领域词语本体的短文本分类方法，该方法以领域高频词作为特征词，经过知网对特征词进行扩展，得到了领域词语本体，并以此为基础进行相似度比较，从而应用于分类。通过对比实验可以证明，准确率、召回率和F1值都有一定的提高。然而由于语言信息的复杂性和灵活性，我们需要不断地探索和完善。

(1) 所建立的本体只是初步原型，要从更多的语义方面进一步优化和扩充词语本体，可以考虑结合其他外部资源如Wordnet和《同义词词林》等进行领域词语本体的扩展；另外，需要进一步人工过滤掉本体中的一些对分类作用不大的虚词而且需要进行特征词的更新和扩充。

(2) 进行相似度计算时，我们通过两个概念的义元之间包含的相同信息量的多少来计算相似度值。例如：两个义元"属性"和"军队"认为对分类的贡献是相同的，但是从分类的角度看，"军队"要比"属性"对分类的贡献更大些。因此，我们可以从节点的深度来考虑给不同层次的义元不同的比重值，并且本文建立的领域词语本体框架为此提供了基础。

参 考 文 献

[1] SebastianiI F. Machine Learning in Automated Text Categorization Consiglio Nazionale delle Ricerche[J]. Italy. ACM Computing Surveys, 2002, 34(1): 1-47

[2] 樊兴华，孙茂松. 一种高性能的两类中文文本分类方法[J]. 计算机学报，2006，29(1)：124-131

[3] Zelikovitz S, Transductive M F. Learning for Short-Text Classification Problem using Latent Semantic Indexing International [J]. Journal of Pattern Recognition and Artificial Intelligence, 2005, 19(2): 143-163

[4] Pu Qiang, Yang Guo Wei. Short-Text Classification Based on ICA and LSA[J]// Proceedings of International Symposium on Neural Networks, 2006(ISNN 2): 265-270

[5] 马后锋，樊兴华. 一种改进的增量贝叶斯分类算法. 仪器仪表学报，2007，28(8III)：312-316

[6] Chen Enhong, Wu Gaofeng. An Ontology Learning Method Enhanced by Frame Semantics [J]// Proceedings of the Seventh IEEE International Symposium on Multimedia. 2005: 374-382

[7] 董振东，董强. 知网. http://www.keenage.com

[8] 郑德权，赵铁军，李生，等. 基于内容的词义本体知识自动获取[A]// 全国第八届计算语言学联合学术会议(JSCL-2005)论文集 [C]. 2005

[9] 刘群，李素建. 基于《知网》的词汇语义相似度计算[A]// 第三届汉语词汇语义学研讨会[C]. 台北，2002

[10] 李峰，李芳. 中文词语语义相似度计算——基于《知网》2000[J]. 中文信息学报，2007，21(3)