

一种基于特征扩展的中文短文本分类方法

王细薇,樊兴华,赵 军

(重庆邮电大学 计算机科学与技术研究所,重庆 400065)

(abeey2007@gmail.com)

摘 要:针对短文本所描述信号弱的特点,提出一种基于特征扩展的中文短文本分类方法。该方法首先利用 FP-Growth 算法挖掘训练集特征项与测试集特征项之间的共现关系,然后用得到的关联规则对短文本测试文档中的概念词语进行特征扩展。同时,引入语义信息并且改进了知网中 DEF 词条的描述能力公式,在此基础上对中文短文本进行分类。实验证明,这种方法具有高的分类性能,其微平均和宏平均值都高于常规的文本分类方法。

关键词:短文本分类;关联规则挖掘;特征扩展

中图分类号: TP311.13; TP391.4 **文献标志码:** A

Method for Chinese short text classification based on feature extension

WANG Xi-wei, FAN Xing-hua, ZHAO Jun

(Institute of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing 400065, China)

Abstract: In this paper, based on the characteristics that short texts describe weak signals, a method based on feature extension (STCFE) was introduced to classify Chinese short texts. In this method, the correlation rules between feature items of training set and testing set were mined by FP-Growth algorithm, and then those rules were applied to extend the features of the testing set. Meanwhile, to classify Chinese short texts effectively, semantic information was introduced and the DEF term formula of words was improved in HowNet. Experimental results show that the proposed method performs well, and its Micro-F₁ and Macro-F₁ are higher than those of conventional approaches.

Key words: short text classification; association rule mining; features extension

0 引言

随着互联网技术与移动通信技术的结合,手机短信、基于互联网和手机短信息的客户评论等中文短文本形式的信息在最近几年进入了爆发式的增长时期,已经成为一种重要的信息传播渠道。这些形式的信息字数不多(大多数为 100 字以内),但是数量非常大。如何从浩瀚的中文短文本中发现有用的信息已经成为信息处理领域亟待解决的关键难点之一,其中,一些实际应用,例如手机短信息过滤和客户评论的分类等,使中文短文本分类问题成为一个重要的研究方向。

目前较为常用的文本分类算法有 Bayes、SVM、KNN、ANN、决策树等^{[1]125}。根据文本特征表示的不同,这些方法可分为两类:一种是传统的基于向量空间模型(Vector Space Model, VSM)的分类,一种是基于语义信息的分类^[2]。一般来说,常规的基于 VSM 的分类方法,没有使用语义信息来改善分类性能;同时,由于短文本数据具有长度短、所描述信号弱的固有特点,在待测文档中出现的词不一定能够在训练集中出现,因而第二种方法亦不能直接应用于短文本分类。

本文结合短文本自身的特点,提出了一种基于特征扩展的中文短文本分类方法 STCFE(Short Text Classification Based on Features Extension)。该方法的基本思路是:一是用 FP-Growth 算法挖掘训练集特征项与测试集特征项之间的共现关系,创建特征共现集并将其用来扩展待测短文本特征;二是在

VSM 中引入知网作为背景知识,将关键词映射到义原,抽取其概念特征,同时改进了知网中 DEF 词条的描述能力公式,在一定程度上解决同义词替换的问题,从而增强短文本特征项的表述能力。

1 基于特征扩展的中文短文本分类方法

基于特征扩展的中文短文本分类流程如图 1 所示。

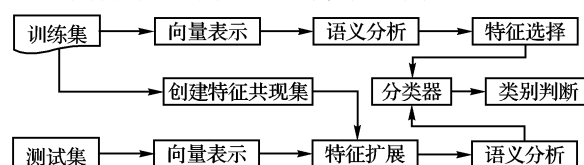


图 1 基于特征扩展的短文本分类基本流程

1.1 关联特征共现集的创建和特征扩展

1.1.1 创建特征共现集

在 VSM 中引入特征项共现,可以对短文本特征进行有效扩展,更好地表示文本的深层语义特征,从而获得更高的短文本分类质量。词共现模型^[3-4]是基于统计方法的自然语言处理研究领域的重要模型之一,它是建立在这样一个基本假设的基础之上:如果在大规模语料中,两个词经常共同出现(共现)在文本的同一窗口单元(如一句话、一个自然段或一篇文本等),则认为这两个词在意义上是相互关联的;并且,共现的概率越高,其相互关联越紧密。我们将词共现模型引入到短文本

收稿日期:2008-09-27;修回日期:2008-11-26。

基金项目:国家自然科学基金资助项目(60703010);重庆市自然科学基金资助项目(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项目(教外司留[2007]1108号);重庆邮电大学科研基金项目(A2006-05)。

作者简介:王细薇(1982-),女,河南许昌人,硕士研究生,主要研究方向:中文信息处理;樊兴华(1972-),男,重庆忠县人,教授,博士,主要研究方向:人工智能、自然语言处理、信息检索;赵军(1971-),男,重庆垫江人,教授,博士,主要研究方向:智能信息处理。

特征扩展中,用来挖掘训练集特征项与测试集特征项之间的共现关系,将超过一定比例的文本中共同出现的特征项看作是相互关联的,并利用 FP-Growth 算法来抽取特征项共现集,只要将特征项作为事务项,将文本看作事务,就可以在给定的最小支持度阈值和最小置信度阈值下发现特征项之间的强关联规则。强关联规则下的特征项则为共现的特征项。

由于每个类的训练样本数不同,为了确保创建的特征共现集能够覆盖到样本中的所有类别,分别计算出每一类别的特征共现集。由于短文本中存在大量的高频但无意义的词语,首先用中文停用词表^[5]过滤这些噪声干扰,只保留汉语句子的核心部分^[6],包括名词、动词、形容词、副词,再将处理结果输入 FP-Growth 算法,按照最小支持度阈值进行第一遍筛选,然后按照最小置信度阈值生成规则,分别得到每一个类别的频繁共现特征集。

1.1.2 关联特征扩展算法描述

输入:特征共现集 I ;

关联规则抽取阈值:最小置信度阈值 C ,最小支持度阈值 S ;待测短文本文档;

输出:短文本特征扩展之后的特征向量空间。

1) 对于待测短文本中的任一个特征项 t_i ,查询特征共现集 I ,如果存在唯一一个规则项 $t_i \rightarrow t_j$,并且当 C 大于设定的阈值时,执行步骤 2)。如果规则项不唯一,则计算 $S * C$,按照 $S * C$ 是最大值的规则进行特征扩展,执行步骤 2)。

2) 提取该联规则右部的特征项 t_j ,如果短文本特征空间列表中不存在 t_j ,执行步骤 3),否则执行步骤 4)。

3) 把 t_j 加入特征空间列表中。

4) 不把 t_j 加入特征空间列表中。

5) 如果短文本特征在特征共现集中找不到匹配方案,执行步骤 6)。

6) 该短文本特征不进行短文本扩展,继续查询下一个短文本特征。

关联规则挖掘算法本身虽然效率不高,但测试时不用每次执行关联规则挖掘算法,只用查询数据库扩展就可以了。由于 t_i 和 t_j 是共现关系,对待测短文本特征的扩展,本质上就是在训练集特征和测试集特征之间建立一种共现关系,将那些相关联的特征词语作为特征补充到短文本中,尽可能地让训练集中出现的词语在测试集中也能出现。

1.2 语义分析方法

经过特征扩展之后的短文本特征向量存在较多的冗余信息,比如同义词的干扰,影响了短文本的特征表达能力。于是,我们利用知网(HowNet)概念词典中的概念树,通过义原在概念树中的位置信息进行概念抽取,并赋予其适当权值来说明其描述能力。具体到每个词,我们计算其 DEF 条目中的权值,并据此决定是将原词选入特征集还是进行概念抽取^[7-8]。期望能提高特征表达准确性,改进分类效率和精度。

《知网》^[9](HowNet)是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。它通过义原的组合来标注各种各样单纯的或复杂的概念,以及各个概念之间、概念的属性和属性之间的关系。相对来说,新词虽然层出不穷,但义原的增加却极少。知网 2002 版共有 1 600 多个义原,其中实体义原(Entity)229 个,事件义原(Event)813 个,次要特征义原(SecondaryFeature)88 个,副词、连词(Attribute/aValue, Quantity/qValue, Syntax)等义原共 482 个。在知网中,词义

就是定义为各种义原的组合,知网 2002 版本中共收录了 6 万余词,每个词的词义都利用义原的组合来解释。

知网中每个词条定义为:W_X = 词语 E_X = 词语例子 G_X = 词语词性 DEF_概念定义。

比如“打”(打篮球,打太极),这个词有一项描述是:

DEF = exercisel 锻炼, sportl 体育

“锻炼”和“体育”就是两个义原。

HowNet 属性特征加权公式^{[8]24}为:

$$k(m) = Wtree_i \cdot \left[\log((Deep_k + 1)/2) + a + \frac{1}{Child_k + b} \right] \quad (1)$$

其中, $k(m)$ 表示义原 m 的权值, i 为 m 所在的义原树的编号, k 为 m 在该义原树中的位置。Wtree 用来调整各棵树的总体重要性,即根节点的权值,由于“Attribute”树、“AttributeValue”树、“Event Role&Features”树、“Quantity”树、“Quantity Value”树、“Syntax”树这六棵树对分类意义较小,文献[8]中它们的 Wtree 的重要度为 0.1;“Entity”树和“Event”树的重要度较高,由于前者大多为名词概念,后者大多为动词概念,因此根据多次实验结果并结合短文本特征特点分析,设前者的 Wtree 为 1.0,后者为 0.65;Deep 表示该义原的高度,Child 为该义原的下位义原的数目,其值从索引表中获得; a, b 为调和因子,用来控制权值范围,防止权值为负值。

由于一个词有很多义原,同时也有很多抽象义原,进行概念抽取之后的表达能力可能不如原词,如果一个词的所有义原表达能力都不强就不进行概念抽取。因此需要一个标准来设置阈值,文献[9]提出了一种计算 DEF 词条能力的描述公式,如式(2)所示:

$$f(c) = \max_{j=0}^m k(c_j) \quad (2)$$

其中, $k(c_j)$ 为词 c 的 DEF 词条中第 j 个义原的权值(由式(1)计算得出), m 为该词 DEF 词条中概念特征的总数目。该公式计算词 c 所有义原的描述能力,并通过该描述能力公式设定阈值来决定选择进行概念抽取或者保留原词。

因为短文本特征有效特征太少,我们对其进行概念抽取的目的就是为了提高其特征描述能力,减少冗余特征,但是如果仍然采用式(2)来设置概念抽取阈值,会把那些对分类有效但是权值较小的义原过滤掉。故本文改进了一个词的 DEF 词条的描述能力公式,如式(3)所示:

$$f(c) = \max_{j=0}^m k(c_j) - \min_{j=0}^m k(c_j) \quad (3)$$

值得注意的是,训练集和待测短文本需要进行相同的语义分析,以保证二者特征的一致性。

2 实验和结果分析

本文所使用的数据集是本项目组收集的 12 个不同领域共 470 252 篇网友评论,其中财经类 35 104 篇、房地产类 28 744 篇、国际新闻类 42 424 篇、国内新闻类 48 288 篇、军事类 49 320 篇、科技类 37 044 篇、女性类 36 032 篇、汽车类 40 372 篇、书评类 39 440 篇、体育类 38 512 篇、游戏类 38 660 篇、娱乐类 36 312 篇。将每类文本集随机地平均分为四份,以其中一份构成测试集,另外三份构成训练集。

对文本分类的性能采用如下四种指标进行评估:宏平均(Macro- F_1)、微平均精确率(Micro-P)、微平均召回率(Micro-R)、微平均 F_1 值(Micro- F_1)。

$$Macro-F_1 = \left(\sum_{i=1}^n F_{1i} \right) / n$$

$$Micro-F_1 = \frac{2 \times Micro-P \times Micro-R}{Micro-P + Micro-R}$$

$$Micro-R = \frac{\sum_{i=1}^n \text{分类器正确分为第 } i \text{ 类的文本数}}{\sum_{i=1}^n \text{测试集中属于第 } i \text{ 类的文本数}}$$

$$Micro-P = \frac{\sum_{i=1}^n \text{分类器正确分为第 } i \text{ 类的文本数}}{\sum_{i=1}^n \text{分类器分为第 } i \text{ 类的文本数}}$$

2.1 两种方法的分类性能比较

常规方法 选用了清华大学开发的 CsegTag3.0 对中文进行分词,并去除停用词,然后用 tf-idf 法将短文本表示成向量,采用 CHI 选择特征方法,以朴素贝叶斯为分类器^[10],测试文档未经过扩展和语义分析,分别选取特征数为 1 000、2 000、3 000、...、10 000,进行 10 轮 12 类短文本分类实验。

STCFE 方法 前面步骤与常规方法相同,只是测试文档经过特征扩展和语义分析,分别选取特征数为 1 000、2 000、3 000、...、10 000,进行 10 轮 12 类短文本分类实验。其中,FP-Growth 的支持度和置信度阈值分别设为 0.01、0.5。

实验结果表明,在特征数为 4 000 时,分类效果就基本趋于稳定,STCFE 方法比常规方法的宏平均 F_1 和微平均 F_1 值都有所提高,均为 6% 左右。

随着 STCFE 方法进行了合理的特征扩展和降低了同义词的噪声干扰,所带来的类别区分信息更加完备,受到噪声数据的干扰也随之减少,STCFE 方法的分类性能提高比较稳定。

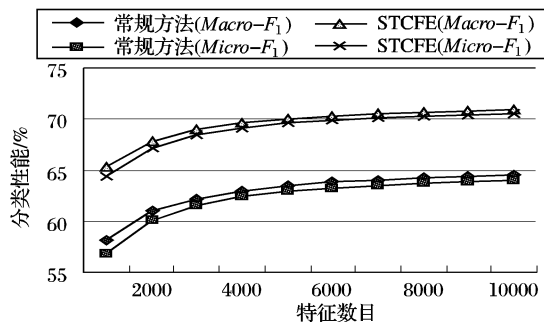


图2 分类性能随特征数目变化曲线

2.2 不同文档类别的性能分析

经过实验,不同的特征数实验效果不一样,但总体说来,STCFE方法在宏平均和微平均指标上都要优于常规方法。典

表1 特征数目为 10 000 时不同类别的性能分析

类别	常规			STCFE		
	精确率/%	召回率/%	F1/%	精确率/%	召回率/%	F1/%
地1类:财经	70.55	62.87	66.49	78.47	70.38	74.21
第2类:房地产	72.73	69.90	71.29	80.25	75.99	78.06
第3类:国际新闻	50.34	47.13	48.68	60.47	56.30	58.31
第4类:国内新闻	59.85	56.94	58.36	67.18	64.34	65.73
第5类:军事	58.90	62.35	60.57	67.86	73.09	70.38
第6类:科技	73.40	57.78	64.66	80.81	68.33	74.05
第7类:女性	59.71	65.28	62.37	64.82	67.91	66.33
第8类:汽车	76.29	72.57	74.38	82.36	79.49	80.90
第9类:书评	77.82	81.55	79.64	80.95	86.05	83.42
第10类:体育	69.76	66.33	68.00	76.94	72.21	74.50
第11类:游戏	65.50	64.94	65.22	68.92	69.58	69.25
第12类:娱乐	46.73	64.66	54.25	49.84	65.28	56.52

型地,当特征数为 10 000 时,其结果如表 1 所示。

从表 1 可以看出,第 11 类和第 12 类提高幅度较小,主要原因是:其本身训练集语料有效特征数目太少而容易分类错误,STCFE 方法对于有效特征太少的数据集判断不够准确,容易受噪声影响。总体来说,除了第 11 类和第 12 类之外的其他 10 个类别数据集的分类性能在经过了特征扩展和语义分析之后其精确率、召回率、F1 值等指标均有大幅度提高。

3 结语

手机短信息过滤和客户评论的分类是中文短文本分类的重要应用之一。针对这些短文本数据的特点提出了一种基于特征扩展的中文短文本分类方法。该方法包括两部分:一是特征共现集的创建。用 FP-Growth 算法挖掘训练集特征项与测试集特征项之间的共现关系,并将其用来扩展待测短文本特征。丰富了短文本特征,克服了由于短文本数据长度短、所描述信号弱的固有缺陷造成的在待测文档中出现的词不一定在训练集中出现的缺点。二是在 VSM 中引入知网作为背景知识,将关键词映射到义原,抽取概念特征,并结合已有的语义分析方法改进了词的 DEF 词条的描述能力公式,在一定程度上解决同义词替换的问题,增强了短文本特征项的表述能力,使用朴素贝叶斯分类方法,实现了对短文本准确高效的分类。研究表明 STCFE 方法在多数短文本分类中是有效可行的,在 12 类共 470 252 篇网友评论构成的语料集上进行多组对比实验,性能和准确度均超过了常规分类方法。

参考文献:

- [1] 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.
- [2] 王永恒,贾焰,杨树强.基于频繁词集聚类的海量短文本分类方法[J].计算机工程与设计,2007,28(8):1744-1780.
- [3] RAK R, STACH W, ZAIAE O R, et al. Considering re-occurring features in associative classifiers [C]// Proceedings of PAKDD, LNCS 3518. Berlin: Springer, 2005: 240-248.
- [4] BAYER T, RENZ I, STEIN M, et al. Domain and language independent feature extraction for statistical text categorization [C]// Proceedings of the Workshop on Language Engineering for Document Analysis and Recognition. Sussex, UK: [s. n.], 1996: 21-32.
- [5] 中文停用词表[EB/OL]. [2008-09-01]. <http://download.cs-dn.net/source>.
- [6] 王元珍,钱铁云,冯小年.基于关联规则挖掘的中文文本自动分类[J].小型微型计算机系统,2005,26(8):1380-1383.
- [7] HE HUI, CHEN BO, XU WEIRAN, et al. Short text feature extraction and clustering for Web topic mining [C]// Proceedings of the Third International Conference on Semantics, Knowledge and Grid. Washington, DC: IEEE Computer Society, 2007: 382-385.
- [8] 董振东,董强.知网及其相关文献[EB/OL]. [2008-09-01]. <http://www.Keenage.com>.
- [9] 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[J].中文信息学报,2005,20(3):22-28.
- [10] 周茜,赵明生,扈雯.中文文本分类中的特征选择研究[J].中文信息学报,2004,18(3):17-23.