

文章编号:1001-9081(2010)10-2621-03

结合语义的特征选择方法

熊忠阳,付玲玲,张玉芳,蒋 健

(重庆大学 计算机学院,重庆 400044)

(mylennfl_00@yahoo.com.cn)

摘 要:传统的基于词频统计的特征选择方法忽略了特征项本身的语义信息,特征项之间存在冗余使得维数有限的特征空间无法容纳更多的对分类有用的特征项。为此,利用《知网》(HowNet)的中英双语知识词典构建“概念—领域”表,对每个词语查询该表,如果在表中,则把该词语映射到“领域”;否则保留原词。这样不仅可以降低低层概念泛化到较高层概念,还能在一定程度上消除特征项之间的冗余,而且从语义上加强它对所在“领域”的分类贡献度。分别应用信息增益和 χ^2 统计利用该方法进行文本分类实验,结果表明该方法可以有效地提高分类准确率。

关键词:文本分类;特征选择;语义;知网

中图分类号: TP391 **文献标志码:** A

Improved feature selection approach combined with semantic

XIONG Zhong-yang, FU Ling-ling, ZHANG Yu-fang, JIANG Jian

(College of Computer Science, Chongqing University, Chongqing 400044, China)

Abstract: The traditional selection methods for text categorization are based on the statistical information of word frequency, which ignores the semantic effect of the words and cannot take more useful features because of the redundancy. A table named "conception-domain" was built based on the semantic dictionary HowNet, which included the word itself and its domain value. If a word from the text was existent in the table, it would be replaced by its domain value with more general meaning. By this way, more semantic information was added to the selected features and the redundancy between features of items could be eliminated to some extent. The experiments were carried out by improved information gain and χ^2 respectively. And the results show that this method has effectively improved the precision of the text categorization.

Key words: text categorization; feature selection; semantic; HowNet

0 引言

文本分类(Text Categorization)是指依据文本的内容,由计算机根据某种自动分类算法,把文本判分为预先定义好的类别^[1],是信息存储、信息检索和信息推送等领域的重要课题。其中,特征空间的高维性是影响文本分类精度和效率的主要原因之一。文本表示是文本分类的前提和基础,其主要工作是确定特征项单位及其权重:选取的特征项单位主要有词、多元词和短语等;权重的计算有词频、文档频率、词频文档倒排频率等。最常用的文本表示模型是向量空间模型(Vector Space Model, VSM)。其基本思想是:把文本表征成由特征项构成的向量空间中的一个点(即一个 n 维向量),通过计算向量之间的距离,来判定文本之间的相似程度。向量空间模型的基本假设是:文档所属类别仅与特征项的频数有关而不考虑特征项所表达的语义信息,这个假设会造成文本信息的丢失;而且,该模型假设特征项之间是正交的,而真实文本的特征项之间存在语义上和结构上的相关性。

1 特征选择的主要方式及存在的问题

VSM把分词后得到的词或词组的集合看做原始特征集,其维数高达几万维甚至几十万维,并且含有大量类别无关项和冗余项。根据John Pierce的理论,用来表示文本的特征项理论上应具有如下特点^[2]:出现频率适中,数量上尽量少,噪

声少,冗余少,与其所属类别语义相关,含义尽量明确。换句话说,只有在类别相关的、没有冗余的特征集上进行训练才能取得更好的效果^[3]。特征选择是根据某种准则从原始特征集中选择部分最有区分类别能力的特征项^[4]。通过特征选择可以留下那些类别区分性强的特征项,降低文本特征向量维数,提高系统速度和分类精度。而且有实验表明,真正具有分类作用的特征项只占总特征项的不到10%^[5]。

特征选择又称独立评估法,是借助某个评估函数对原始特征项逐一评分,根据分值大小选取分值最高的前 n 个特征项作为原始特征集 S 的一个子集 S' ^[6]。目前,常用的特征选择方法主要有基于特征频数统计的文档频率法(Document Frequency, DF)、互信息(Mutual Information, MI)、信息增益(Information Gain, IG)、期望交叉熵(Expected Cross Entropy, CE)、 χ^2 统计(Chi-square, CHI)等。已有的实验表明:IG和 χ^2 统计是最有效的两种特征选择方法^[7]。

但是这些方法的指导思想是基于特征项频数或特征项与类别之间的相关性的统计信息来选取表示文本的特征项集合。这种基于词形统计的特征选择方式存在的问题主要有:忽视了词语与类别之间的语义联系;由于同义词、近义词、相关词等的大量存在,被挑选出来的特征项之间存在语义上的冗余,这会使得在特征空间维数有限的情况下,一些对文本分类含有有用信息的特征项却由于其分值略低于前面的一些冗余特征项的分值而没有被选择出来,这样选择得到的特征子

收稿日期:2010-04-06;修回日期:2010-06-09。

基金项目:中国博士后科学基金资助项目(20070420711);重庆市科委基金资助项目(CSTC2008BB2191)。

作者简介:熊忠阳(1962-),男,重庆人,教授,博士生导师,主要研究方向:数据挖掘、网络技术、并行计算;付玲玲(1983-),女,四川营山人,硕士研究生,主要研究方向:数据挖掘、自然语言处理;张玉芳(1965-),女,上海人,副教授,主要研究方向:数据挖掘、网络入侵检测、并行计算;蒋健(1986-),男,江西人,硕士研究生,主要研究方向:机器学习、数据挖掘。

集 S' 在类别界定作用的意义上并不是 S 的最佳选择^[6]。针对这两个问题目前的解决方法有:1)特征抽取,如潜在语义分析法、主要成分分析法以及文献[6]提出的二次特征选择法等;2)结合上下文进行语义排歧来处理同义词近义词等,如文献[8];3)利用语义词典计算词语间的相似度来消除冗余特征项,如文献[9]。方法1)是基于统计思想的降维方法,没有真正考虑特征项间的语义联系;方法2)和3)通过处理特征项间的横向联系消除冗余,但是计算公式中参数复杂且难以确定,导致计算开销大而效果并不太好。为此,本文利用《知网》提出一种方法,通过处理特征项之间的纵向联系来解决上述两个问题。

2 结合语义的特征选择方法

2.1 《知网》简介

《知网》(HowNet)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^[10]。下面先介绍本文用到的《知网》中的几个重要概念。

概念 是对词语语义的一种描述,一个词语的概念(也称为义项)及其描述形成一个记录(记录格式后面有说明)。由于词语存在多义,因此一个词语可能表达为一个或几个概念,每个概念形成一个记录。

义原 《知网》用一种“知识表示语言(Knowledge Dictionary Mark-up Language, KDML)”来描述概念,该语言所用的特定词汇称为义原。义原是描述一个概念的最基本的、不易于再分割的意义的最小单位,《知网》试图用一系列义原来对每一个概念进行描述。比如“education|教育”就是一个义原。

上下位关系 给定概念 C_1 和 C_2 ,若 C_2 的外延包含 C_1 的外延,则 C_1 和 C_2 具有上下位关系,称 C_2 为 C_1 的上位概念, C_1 为 C_2 的下位概念^[11]。《知网》中的义原之间存在复杂的关系,共有八种,其中上下位关系是最主要的关系。

本文主要使用了《知网》的中英双语知识词典。词典中每个词语用一条或多条记录来表示,记录样式如下:

```
NO. = 025286
W_C = 大学
G_C = V[da4 xue2]
S_C =
E_C =
W_E = university
G_E = N
S_E =
E_E =
DEF = { InstitutePlace|场所:
domain = { education|教育},
modifier = { HighRank|高等},
{ study|学习: location = { ~ }},
{ teach|教: location = { ~ }}}
```

其中:NO.为概念编号;W_C、G_C、E_C分别是对应的汉语词语、词性和例子;W_E、G_E、E_E分别是对应的英语的词语、词性和例子;DEF是概念的定义,表达了这个概念的语义信息。本文用到了DEF中的domain这个动态特征的值所包含的领域信息。domain的值可能是与进行文本分类相对应的领域信息,如计算机、经济、政治、体育等,也可能是该概念的上位概念,比如在“棒球场”这个概念的DEF中domain的值是“棒球”。

2.2 构建“概念—领域”表

为了便于处理文本中词语与类别之间的语义联系以及词

语的上下位关系,先根据《知网》构建一个“概念—领域”表,构建规则如下。

- 1)只处理具有一个DEF的词语,对多义词不作处理;
- 2)从每个记录中取汉语词语作为“概念”值,取DEF中第一个domain的值作为“领域”值;
- 3)“概念—领域”表中无重复记录。

对于多义词,即有多个DEF定义,由于语义排歧的难度和开销都很大,而且需要大量地考虑词语的上下文,所以在构建“概念—领域”表时没有处理多义词。

2.3 改进的特征选择

利用构建好的“概念—领域”表进行特征选择。为了方便叙述,采用 $S(t)$ 表示原始的特征评估函数, $S_1(t)$ 为进行概念映射之后的特征评估函数, t 是特征项。对文本预处理、分词之后,把每个词语 t 都进行一次到“领域”的映射操作:

- 1)如果词语 t 在“概念—领域”表中的“领域”值 D 等于该词语当前所在文档 T 的所属类别 C_i ,则该词语 t 对类别 C_i 的类别区分度放大 $r(r > 1)$ 倍,改进后的特征评估函数如下:

$$S_1(t) = S(t) \times r \quad (1)$$

参数 r 的最佳取值通过实验获得。

- 2)如果词语 t 在“概念—领域”表中的“领域”值 D 不等于训练集中的任何一个类别 C_i ,则该词语的“领域”值 D 表示的并非类别信息,而是词语 t 的上位词,于是将词语 t 替换成它的上位词 D 。如果 t 的上位词 D 在原始特征集中不存在则加入 D ,并将词语 t 的词频统计值增加给上位词 D ,然后从原始特征集中移除词语 t 。比如,在“概念—领域”表中“棒球场”的“领域”值是“棒球”,于是把“棒球场”替换成“棒球”,假设“棒球场”的词频统计值为 t_c ,“棒球”的词频统计值为 D_c ,则替换之后“棒球”的词频统计值为 $D_c + t_c$ 。然后采用原始特征评估函数计算替换后的上位词 D 的函数值为 $S(D)$ 。

- 3)如果词语 t 在“概念—领域”表中不存在,则保持这些词及其统计值,采用 $S(t)$ 计算它们的函数值。

这样处理可以解决以下问题:①尽管用统计方法也可以发现某些词经常出现在某一类别的文章中,但是步骤1)从语义的角度进一步加强了类别核心词在分类时的作用,既利用了词语统计信息又结合语义信息;另外,步骤1)还可以放大部分对分类有用但词频统计值略低的特征项的作用,增加其被选入特征子集 S' 的几率。②一部分词语不会直接映射到文章所属的类别而是映射到它的上位词(上位概念),也就是上面的步骤2),它实际上是处理了一部分同义词、近义词和相关词,将较低层概念泛化到较高层概念,在一定程度上消除了特征项之间的冗余,因为高层概念能提供数据“更清晰”的概括,揭示更一般的概念^[12]。比如,查询知识库发现“棒球场”、“棒球队”、“棒球帽”、“棒球赛”、“棒球队员”、“棒球手套”这些词的DEF中domain的值都是“{baseball|棒球}”,而这些词语在语义上是相关的,都与“棒球”相关,于是把这6个词都映射成“棒球”,能减少特征项间冗余又增强上位词的重要性,而且“棒球”是比“棒球场”、“棒球队”等更一般的概念,携带了更多的信息。

2.4 分类阶段的改进

对于测试文档,预处理之后将每篇文档映射到特征空间生成文档向量,传统的做法是进行“词形”的映射,即只有词语 t 在特征空间中存在才能映射。本文利用“概念—领域”表对这个过程进行了改进,增加对概念的映射,如果词语 t 在特征空间中不存在,但它的上位词 D (上位概念)在特征空间存在,就把 t 映射给它的上位词 D (上位概念)并把词语 t 的词频统计值计入 D 的词频统计值。这样可以在不增加空间维数的前提下减少文档语义信息的丢失。

整个系统流程如图1所示。

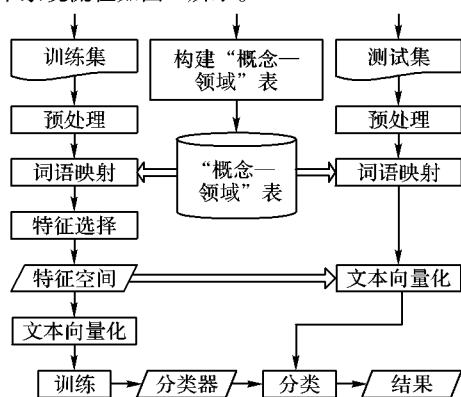


图1 改进特征选择后的系统流程

3 实验及结果分析

把上述特征选择的思路分别应用于信息增益IG和 χ^2 统计,并在文本分类方面进行了实验。实验数据采用复旦大学李荣陆教授提供文本数据集^[13],从中挑选了7个类,共2151篇文章采用四分交叉法进行实验,将2151篇文章平均分为4份,每次取3份为训练集,1份为测试集,循环4次;取4次的平均值作为测试结果。文本分类器采用K-最近邻分类算法(K-Nearest-Neighbor, KNN),评价指标采用查准率、查全率。

查准率 = 分类正确的文本数/实际分类文本数

查全率 = 分类正确的文本数/应有文本数

为了与传统的特征选择方法的分类效果进行比较,用KNN分类器对未改进的特征选择方法也进行了实验。分类结果统计如表1、2所示。

表1 改进前后的IG方法在文本分类实验中的结果统计

特征选择	查全率/%		查准率/%	
	改进前	改进后	改进前	改进后
教育	90.2	94.0	94.5	98.1
计算机	76.8	83.0	94.0	100.0
政治	95.0	97.1	91.3	93.0
经济	92.0	95.4	69.0	73.6
军事	75.7	80.5	93.4	96.0
体育	89.0	93.4	90.6	95.4
医疗	79.5	83.9	94.8	98.0
宏平均	85.5	89.7	89.6	93.4

表2 改进前后的 χ^2 统计方法在文本分类实验中的结果统计

特征选择	查全率/%		查准率/%	
	改进前	改进后	改进前	改进后
教育	91.4	97.1	95.1	98.6
计算机	88.0	94.0	96.0	100.0
政治	94.5	96.2	89.5	93.6
经济	93.0	95.4	65.0	76.4
军事	76.8	81.3	94.6	96.6
体育	90.8	93.4	94.2	97.5
医疗	75.1	81.5	97.8	100.0
宏平均	87.1	91.3	90.3	94.7

分析表中的数据发现:对特征项引入语义信息后,分类的查准率和查全率都有较大的提高,因此本文提出的结合语义的特征选择方式总体效果优于传统的特征选择方式;同时,与现有的利用语义进行特征选择的其他方法(如语义排歧、词语相似度计算等)相比,本文的方法抓住知识词典中的与分类有关的主要信息,忽略次要信息减少了计算复杂度。

4 结语

特征选择是文本处理所必须面对的主要问题之一。特征选择方法的好坏直接影响到特征项的质量和分类器的准确率。本文提出的引入《知网》结合语义和统计两方面信息选择出类别信息较强特征项,更好地表达文章的主题;同时,又能够在一定程度上消除特征项之间的冗余,使得维数有限的特征空间里包含更多的对分类有用的项。如何更有效地利用本体进行文本分类将是我们的下一步的研究工作。

参考文献:

- [1] 周茜,赵明生,扈旻.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.
- [2] 汪建华.中文文本分类技术研究[D].长春:吉林大学,计算机科学与技术学院,2007.
- [3] YU LEI, LIU HUAN. Efficient feature selection via analysis of relevance and redundancy [J]. Journal of Machine Learning Research, 2004, 5: 1205-1224.
- [4] 徐燕,李锦涛,王斌,等.基于区分类别能力的高性能特征选择方法[J].软件学报,2008,19(1):82-89.
- [5] 刘丽珍,宋瀚涛.文本分类中的特征选取[J].计算机工程,2004,30(4):14-16.
- [6] 刘海峰,王元元,姚泽清,等.文本分两维:一种基于选择的二次特征降维方法[J].情报学报,2009,28(1):23-27.
- [7] YANG YIMING, PEDERSEN J O. A comparative study on feature selection in text categorization [C]// ICML '97: Proceedings of the Fourteenth International Conference on Machine Learning, San Francisco, CA: Morgan Kaufmann Publishers, 1997: 412-420.
- [8] 蒋敏梅.基于概念的文本分类研究[D].北京:北京交通大学,2008.
- [9] 赵长伟,孙素环,李晓培.基于语义相似度的文本表示降维方法[J].河南科技大学学报:自然科学版,2008,29(5):36-39.
- [10] 知网[EB/OL]. [2009-12-06]. <http://www.keenage.com/html/c-index.html>.
- [11] 刘磊,曹存根,张春霞,等.概念空间中上下位关系的意义识别研究[J].计算机科学,2009,32(8):1651-1661.
- [12] 贾焰,王永恒,杨树强.基于本体论的文本挖掘技术综述[J].计算机应用,2006,26(9):2013-2015.
- [13] 李荣陆.文本数据集[EB/OL]. [2009-12-08]. http://www.nlp.org.cn/categories/default.php?cat_id=16.

征订通知

本刊以应用技术为主,内容丰富多彩,现审稿周期为2个月,发表周期为6个月。欢迎投稿,欢迎订阅。全国各地邮局均可订阅,也可直接从编辑部订阅。

邮发代号:62-110

定价:28元/册,全年336元/12期。

通信地址:成都市237信箱《计算机应用》编辑部

邮政编码:610041 联系人:雍平

电话:028-85224283(803)

传真:028-85222239(816)

开户名称:计算机应用杂志社

开户银行:交行成都市分行科分院支行

账号:511609017018001969114

作者优惠订刊详情请见:

<http://www.computerapplications.com.cn/CN/column/item130.shtml>

本刊编辑部

2010年10月