基于语义距离的文本分类方法

张培颖,王雷全
(中国石油大学(华东) 计算机与通信工程学院,山东 青岛 266580)

摘 要:文本分类是解决网络信息过载的关键技术之一。传统的文本分类方法大多只考虑文本中词语的统计词频等特征,忽略了文本的语义信息,导致文本分类精度不高。针对这种问题,提出了一种基于语义距离的文本分类方法,该方法首先根据 CHI 方法进行文本特征选择,然后利用语义距离计算代表类别的特征向量集合,最后通过计算文本特征向量和类别特征向量之间的语义距离来确定文本类别。实验结果表明,该方法与其他方法相比,把文本的语义信息考虑在内,在进行文本分类方面具有较高的准确率。

关键词:文本分类;语义距离;特征选择;特征向量
中图分类号:TP391 **文献标识码**: A **文章编号**:1673-629X(2013)01-0128-03
doi:10.3969/j.jssn.1673-629X.2013.01.032

Text Classification Method Based on Semantic Distance

ZHANG Pei-ying, WANG Lei-quan
(College of Computer & Communication Engineering, University of Petroleum (East China), Qingdao 266580, China)

Abstract: Text classification is one of the key technologier solving network information overload. The traditional text classification methods only consider mostly the statistical word frequency in the text, ignoring the text semantic information, leading to text classification accuracy is not high. Aiming to this problem, propose a text classification method based on semantic distance, this method first take the text feature selection according to CHI method, and then by using semantic distance calculate feature vector set on behalf of the category, finally through the calculation of the distance between text feature vector and category feature vector determine the text category. The experimental results show that this method, compared with the existing methods, take the semantic information of the text into consideration, has higher accuracy in the text classification.

Key words: text classification; semantic distance; features selection; feature vector

0 引 言

随着网络信息技术的高速发展,用户所面临的文本信息量急剧膨胀,面对网络上众多的文本信息,用户查找所关心的信息变得越来越困难,文本自动分类技术也就越来越受到研究学者的普遍关注。文本分类作为有效地组织和管理这些海量文本信息的一种手段,目前在个性化信息检索、垃圾邮件的过滤、面向特定领域的主题搜索、个性化信息推送等领域得到了广泛的应用。文本分类的任务就是把待分类的文本文档映射到提前定义好的文本类别的过程。

经过国内外学者的不断研究,已经在文本分类领域存在了许多种不同的方法[1~4]。常见的文本分类方法有 KNN 分类方法[5],Naïve Bayes 分类方法和 Support Vector Machine (SVM)分类方法[6,7],这些分类方法都是采用向量空间模型(VSM)来表示文本的特征,向量空间模型把文档看作是以关键词的权重为分量的一组正交向量所构成的向量空间,每篇文档被表示为其中的一个关键词特征向量,这样每篇文档就被表示为向量空间中的一个点,于是文档集合中所有文档的匹配问题就转换为向量空间中的向量匹配问题[8]。

基于关键词的分类方法采用的是向量空间模型表示法,该模型把文本表示为许多词的集合(bag of word),而没有考虑词语之间的顺序信息和语义信息。文中提出的基于语义距离的文本分类方法,主要从挖掘文本语义信息的角度,利用文本特征向量和类别特征向量之间的语义距离进行分类。

收稿日期:2012-05-14;**修回日期**:2012-08-20
基金项目:中央高校基本科研专项资金(09CX04031A);中国石油大学(华东)计算机与通信工程学院青年教师创新基金(08120907)
作者简介:张培颖(1981-),男,硕士,讲师,主要研究领域为自然语言理解、信息检索。

1 相关工作

1.1 向量空间模型

向量空间模型是由 Salton 于 20 世纪 60 年代首先

提出的,到目前为止仍然是中文信息处理技术研究的基础,也是应用最多最普遍的模型。

向量空间模型的核心思想就是把中文语言表示的文本转换为数字表示的向量形式。例如文本文档 D 可以表示为:$(w_1, w_2, \cdots, w_n)$,其中 w_i 表示第 i 个特征项的权重系数;n 代表文本文档 D 中特征项的个数。特征项指的是在文本文档中能代表该文档的字、词或者短语,一般都是由经过中文分词之后,并过滤掉停用词之后的名词、动词等组成。特征项的权重系数主要反映该特征项能够代表文本文档的程度,一般都是选择某种算法来衡量特征项的重要程度,比较常用的算法主要有:布尔表示法、TF-IDF 法、TF-CRF 法等。布尔表示法最为简单,即文本文档中包含该词,那么该词的权重系数就为 1,否则权重系数就为 0。但是这种方法显然过于简单,无法体现特定的词语在文本文档中的重要程度,因此被更加精确的计算方法所代替,目前普遍采用的是 TF-IDF 法,计算公式如下:

$$w(t,d) = \frac{tf(t,d) \times \log(\frac{N_d}{df})}{\sqrt{\sum_{i=1}^{N_t} [tf_i(t,d) \times \log(\frac{N_d}{df_i})]^2}} \tag{1}$$

其中,$w(t,d)$ 代表该特征词 t 在文本文档 d 中的权重,N_d 代表文档的总数,N_t 代表特征词的总数,$tf(t,d)$ 代表该特征词 t 在文本文档 d 中的词频,df 代表训练文档集中出现该特征词 t 的文档总数。

1.2　语义距离计算方法

词语之间语义距离的计算方法主要分为两类:一类是基于统计的方法;另一类是基于世界知识的方法。基于统计的方法主要是利用大规模语料库进行统计,计算结果依赖于语料库的大小,现实世界中很难获取大规模的语料库,因此大多数的研究都采用基于世界知识的计算方法。基于世界知识的计算方法主要利用一个分类体系的数据库,利用分类树中节点之间的路径来衡量节点之间的语义距离。

刘群,李素建[9]利用《HowNet》[10]来计算词语之间的相似度,主要是根据义原之间的上下位关系进行计算距离的,词语之间的相似度是通过综合考虑组成词语的义原,对义原之间的相似度进行加权处理便构成了词语之间的相似度。Wu Zhi-biao 和 Palmer Martha 利用 WordNet 来计算词语之间的相似度,通过词网中两个概念节点及其最小公共父节点在层次目录树中的深度来计算词语之间的相似度[11],计算公式如下式所示:

$$\mathrm{sim}(c_1,c_2) = \frac{2 \times \mathrm{depth}(lcs)}{\mathrm{len}(c_1,lcs) + \mathrm{len}(c_2,lcs) + 2 \times \mathrm{depth}(lcs)} \tag{2}$$

其中:lcs 代表节点 c_1 和 c_2 在 WordNet 层次结构中的最小公共父节点,$\mathrm{depth}(lcs)$ 代表节点 lcs 在层次结构中的深度,$\mathrm{len}(c_1,lcs)$ 或 $\mathrm{len}(c_2,lcs)$ 分别代表节点 c_1 或 c_2 到 lcs 的距离。

国内的学者主要是利用知网进行语义距离计算的研究,比较有代表性的主要是刘群提出的词语相似度计算方法,后来很多算法都是在此基础上进行改进的。因此文中采用刘群的计算方法来计算词语之间的相似度,还利用《同义词词林扩展版》进行识别同义词。

2　基于语义距离的文本分类方法

2.1　词语之间的语义距离

词语之间的语义距离计算要利用一定的语义知识资源作为基础,这里采用《HowNet》[10]和哈尔滨工业大学信息检索实验室的《同义词词林扩展版》作为系统的语义知识资源来计算词语之间的相似度。HowNet 是由董振东教授历经十余年开发的一个语义知识资源库,词语表示为若干个概念,概念是由若干个义原通过“知识描述语言”来进行组织的。HowNet 中有两个关键的概念:“概念”和“义原”。概念就是常说的词语,义原是用来对概念进行描述的最小的语义单元。概念和义原之间的关系是采用被称为“知识描述语言”来进行组织的。

词语 W_1 和 W_2,W_1 有 n 个概念:$S_{11}, S_{12}, \cdots, S_{1n}$,$W_2$ 有 m 个概念:$S_{21}, S_{22}, \cdots, S_{2m}$。这里规定:$W_1$ 和 W_2 的相似度取各个概念的相似度的最大值,即:

$$\mathrm{sim}(W_1,W_2) = \max_{i=1\cdots n, j=1\cdots m} \mathrm{sim}(S_{1i},S_{2j}) \tag{3}$$

“义原”之间的相似度计算公式如下式所示:

$$\mathrm{sim}(p_1,p_2) = \frac{\alpha}{\alpha + d} \tag{4}$$

式中 p_1 和 p_2 代表两个义原(primitive),d 代表 p_1 和 p_2 在义原层次体系树中的路径长度,是一个正整数。α 是一个可调节的参数。

词语之间的相似度计算方法:首先查找同义词表,如果为同义词,则词语之间的相似度为 1;否则,词语之间的相似度采用文献[9]中的计算方法,计算公式如下式所示:

$$\mathrm{sim}(S_1,S_2) = \sum_{i=1}^{4} \beta_i \prod_{j=1}^{i} \mathrm{sim}_j(S_1,S_2) \tag{5}$$

其中:$\beta_i (1 \leqslant i \leqslant 4)$ 是可调节参数,且有:$\beta_1 + \beta_2 + \beta_3 + \beta_4 = 1$,$\beta_1 \geqslant \beta_2 \geqslant \beta_3 \geqslant \beta_4$。

2.2　文本特征向量抽取

把文本表示为特征向量进行处理时,会遇到特征的高维性和文本向量数据的稀疏性问题。如何降低特征向量空间的维数成为提高文本分类效率的关键。降低维数通常采用特征选择和特征抽取两种方法[12]。

特征选择主要是根据某种计算准则从众多原始特征空间中选取最能代表类别统计特性的特征,特征选择不改变原始特征空间的特征词语。常用的特征选择方法有文档频率法、特征熵法、互信息法、CHI 统计法、期望交叉熵等。

特征抽取是利用特征项之间的语义联系、特征反映类别的统计特性、类别之间的离散程度等方面来考虑对原始特征空间的一种压缩处理。特征抽取包含高维矩阵分解的困难。常用的特征抽取方法主要包括:LSI 方法、矩阵分解法等。

文中采用特征选择方法进行特征的降维处理,文本特征向量的选择是根据某种算法,计算文本中每个词汇的权重,选取权重较大的若干个词语作为文本的特征向量。这里采用效果较好的 CHI 方法进行特征向量的选取,对于类别 C,文本特征词条 t 的 CHI 值定义如下式所示:

$$\mathrm{CHI}(t,C)=\frac{[P(t,C)P(\bar{t},\bar{C})-P(t,\bar{C})P(\bar{t},C)]^2}{P(t)P(\bar{t})P(C)P(\bar{C})} \quad (6)$$

其中:$\mathrm{CHI}(t,C)$ 表示词条 t 的 CHI 值,$P(t,C)$ 为训练集中出现词条并且属于类别 C 的样本数除以训练集的样本数,$P(\bar{t},\bar{C})$ 为训练集中不出现词条 t 并且不属于类别 C 的样本数除以训练集的样本数,$P(t,\bar{C})$ 为训练集中出现词条 t 并且不属于类别 C 的样本数除以训练集的样本数,$P(\bar{t},C)$ 为训练集中不出现词条 t 并且属于类别 C 的样本数除以训练集的样本数。

2.3 特征向量集合之间的距离

假设两个特征向量集合 $\boldsymbol{V}_1=\{w_1,w_2,\cdots,w_m\}$,$\boldsymbol{V}_2=\{q_1,q_2,\cdots,q_n\}$,基于词语之间相似度,可以定义两个特征向量集合之间的语义距离如下式所示:

$$\mathrm{Sim}(\boldsymbol{V}_1,\boldsymbol{V}_2)=\frac{\sum_{1\leqslant i\leqslant m}\max_{1\leqslant j\leqslant n}\mathrm{Sim}(w_i,q_j)+\sum_{1\leqslant j\leqslant n}\max_{1\leqslant i\leqslant m}\mathrm{Sim}(w_j,q_i)}{|\boldsymbol{V}_1|+|\boldsymbol{V}_2|} \quad (7)$$

其中:$|\boldsymbol{V}_1|$ 和 $|\boldsymbol{V}_2|$ 分别表示两个特征向量集合中元素的个数。

2.4 文本分类的实现过程

基于语义距离的文本分类方法的实现流程图如图 1 所示:

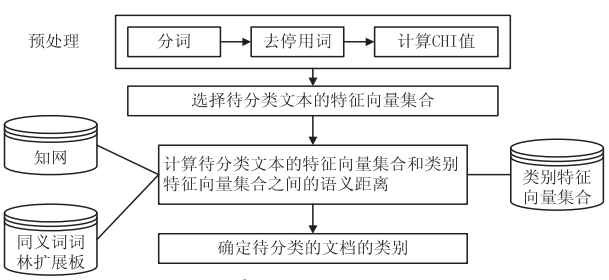

图 1 文本分类的处理流程图

输入:待分类的文本

输出:文本的类别

具体的实现过程如下所示:

1. 把待分类的文本进行预处理,主要包括:分词,去停用词,计算每个词语的 CHI 值;

2. 选取 CHI 值最大的前 20 个词语作为待分类文本的特征向量集合;

3. 计算训练集中每个类别的所有文本的特征向量和类别之间的语义距离,选取距离最近的 20 个词语作为类别的特征向量集合;

4. 计算待分类的文本特征向量集合和每个类别特征向量集合之间的语义距离;

5. 选择距离最近的类别作为待分类文本的类别。

3 实验结果及分析

3.1 性能评价

文本分类中通常采用的性能评估指标有:准确率、召回率和 F1 值。准确率是所有判断的文本中与人工分类结果相符的文本所占的比率,计算公式如下:

$$\text{准确率}=\frac{\text{正确分类文本数}}{\text{实际分类文本数}}$$

召回率是人工分类结果应用的文本中与分类系统相符的文本所占的比率,计算公式如下:

$$\text{召回率}=\frac{\text{正确分类文本数}}{\text{应用文本数}}$$

$F1$ 值是将准确率和召回率综合起来的一个指标。计算公式如下:

$$F1=\frac{\text{准确率}\times\text{召回率}\times 2}{\text{准确率}+\text{召回率}}$$

3.2 实验结果及分析

本实验选取了目前常用的 KNN 分类方法、SVM 分类方法和基于语义距离的分类方法作比较,实验结果如表 1 所示:

表 1 3 种文本分类方法的分类效果

分类方法	准确率(%)	召回率(%)	$F1$ 值
KNN 方法	90.13	85.72	87.87
SVM 方法	90.34	86.74	88.50
语义距离方法	90.42	87.52	88.95

从上表可以看出,SVM 分类方法优于 KNN 分类方法,语义距离分类方法优于 SVM 分类方法。实验结果表明,基于语义距离的文本分类方法在一定程度上把文本语义信息考虑在内,分类效果优于目前常用的分类方法。

4 结束语

文本分类在中文信息处理、文本信息的组织与管

(下转第 134 页)

密码协议是信息和通讯安全技术的重要内容,而密码的应用已经从军事、政府和外交迅速扩展到了工业、商业和金融等领域,而对于密码协议的研究也越来越重要。而在文中,提出了一种抛掷硬币协议,这种协议比较简单便捷地解决硬币抛掷问题的方法——基于雅克比符号的硬币抛掷问题的解决方法,这种方法能够使得不相互信任的双方通过抛掷硬币的方式,对于争执的问题达成共识。硬币抛掷问题的解决有着很大的意义[11,12],比如,通过硬币抛掷协议,可以使得 Alice 和 Bob 产生随机会话密钥,以便双方都不能影响密钥产生的结果,从而完成 Alice 和 Bob 在网络上交换邮件消息或其他通信。

参考文献:

[1] Ambainis A, Buhrman H, Dodis Y, et al. Multiparty quantum coin flipping[C]//Proceeding of the 19th Annual IEEE Conference on Computational Complexity. [s. l.]:[s. n.],2004:250-259.

[2] Saks M. A robust noncryptographic protocol for collective coin flipping[J]. SIAM Journal on Discrete Mathematics, 1989, 2(2):240-244.

[3] Ben O M, Linial N. Collective coin flapping[J]. Advances in Computing Research: Randomness and Computation, 1989(5):91-115.

[4] Alon N, Naor M. Coin-flipping games immune against linear-sized coalitions[J]. SIAM Journal on Computing, 1993, 22(2):403-417.

[5] 吴晓平. 信息安全数学基础[M]. 北京:国防工业出版社,2009:66-69.

[6] Ambainis A. A new protocol and lower bounds for quantum coin flipping[J]. Journal of Computer System Sciences, 2004, 68(2):398-416.

[7] Gordon S D, Hazay C, Katz J, et al. Complete fairness in secure two-party computation[C]//Proceedings of the 40th Annual ACM Symposium on Theory of Computing. [s. l.]:[s. n.],2008:413-422.

[8] Lindell Y. Parallel coin-tossing and constant-round secure two-party computation[J]. Journal of Cryptology, 2003, 16(3):143-184.

[9] Scheneier B. 应用密码学-协议、算法与 C 源程序[M]. 吴世忠,祝世雄,张文郑,等译. 北京:机械工业出版社,2010:62-65.

[10] 李顺东. 现代密码学:理论、方法与研究前沿[M]. 北京:科学出版社,2008:143-144.

[11] 付潇潇,王世民. 基于 RSA 加密算法的扑克游戏[J]. 北京工商大学学报(自然科学版),2007(5):60-63.

[12] 佘 堃,沈 仟,周明天. 背包问题在硬币抛掷协议上的研究[J]. 电子科技大学学报,2003,32(4):417-419.

(上接第 130 页)

理、垃圾邮件过滤等领域有着广泛的应用,是解决网络信息过载的有效途径之一。运用语义的知识来进行文本分类是目前国内外学者研究的热点。文中提出了一种基于语义距离的文本分类方法,首先利用 CHI 特征选择方法进行文本特征选择,然后利用词语之间的距离计算代表类别的特征向量集合,最后通过计算文本特征向量和类别特征向量之间的语义距离来确定文本类别[13,14]。实验结果取得了较高的准确率,但该方法受词语相似度计算结果的影响,如果能进一步提高词语之间的相似度计算的准确率,将得到更好的结果。

参考文献:

[1] Zhang W, Taketoshi Y, Tang X J. Text Classification Based on Multi-word with Support Vector Machine[J]. Knowledge-based Systems, 2008, 21(8):879-886.

[2] Chen Y T, Chen M C. Using Chi-square Statistics to Measure Similarities for Text Categorization[J]. Expert Systems with Applications, 2011, 38(40):3085-3090.

[3] Wang Jun, Zhou Yiming. A Novel Text Representation Model for Text Classification[C]//First International Conference on Intelligent Networks and Intelligent Systems. [s. l.]:[s. n.], 2008:702-705.

[4] 杨金柱,刘金岭. 基于词语上下文的文本分类研究[J]. 计算机技术与发展,2011,21(8):145-148.

[5] 鲁 婷,王 浩,姚洪亮. 一种基于中心文档的 KNN 中文文本分类算法[J]. 计算机工程与应用,2011,47(2):127-130.

[6] 张 苗,张德贤. 多类支持向量机文本分类方法[J]. 计算机技术与发展,2008,18(3):139-141.

[7] 姜 鹤,陈丽亚. SVM 文本分类中一种新的特征提取方法[J]. 计算机技术与发展,2010,20(3):17-19.

[8] 林 伟,孟凡荣,王志晓. 基于概念特征的语义文本分类[J]. 计算机工程与应用,2011,47(28):139-142.

[9] 刘 群,李素建. 基于知网的词汇语义相似度计算[C]//第三届汉语词汇语义学研讨会. 台北:出版者不详,2002.

[10] 董振东. 知网[DB/OL]. 2012. http://www. keenage. com.

[11] Wu Zhibiao, Martha P. Verb Semantics and Lexical Selection[C]//Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics. New Mexico: Association for Computational Linguistics, 1994:133-138.

[12] 胡 涛,刘怀亮. 中文文本分类中一种基于语义的特征降维方法[J]. 现代情报,2011,31(11):46-50.

[13] 张培颖. 基于句子特征和语义距离的文本摘要技术[J]. 微计算机应用,2009(7):84-89.

[14] 宋 玲,马 军,连 莉. 文档相似度综合计算研究[J]. 计算机工程与应用,2006,42(30):160-163.

基于语义距离的文本分类方法

作者： 张培颖， 王雷全
作者单位： 中国石油大学华东 计算机与通信工程学院,山东 青岛 266580
刊名： 计算机技术与发展
英文刊名： Computer Technology and Development
年，卷(期)： 2013(1)