一种面向电子邮件分类的特征值处理方法

邹 娟,周经野,邓 成,陈 静
ZOU Juan,ZHOU Jing-ye,DENG Cheng,CHEN Jing

湘潭大学 信息工程学院,湖南 湘潭 411105
Information Engineering College,Xiangtan University,Xiangtan,Hunan 411105,China
E-mail:zoujuan@xtu.edu.cn

ZOU Juan,ZHOU Jing-ye,DENG Cheng,et al.Characteristic value extractive method for e-mail categorization.Computer Engineering and Applications,2007,43(18):177-180.

Abstract: Based on the e-mail,we propose a method of characteristic value extraction in this paper.We process synonym and polysemant that use of the fuzzy theory.So the characteristic value that using the method of characteristic value extraction in this paper can the better denote text characteristic.Finally,we present the results of the experiments comparing with other characteristic value extraction method and the results of applying the method in the different classified algorithm,which illustrate that the method in this paper improve the correct rate of e-mail categorization and reduce the dimensions effectively.
Key words: e-mail;text categorization;characteristic value extraction

摘 要:利用电子邮件的特点提出了一种面向电子邮件分类处理的特征值提取方法。本方法根据电子邮件文法随意性的特点,利用模糊集合对其同义词和多义现象都进行了处理,使得所得到的特征值能更好的契合文本的特点。通过与其它特征值提取方法的比较实验,以及在不同分类算法中应用实验结果都证明文中提出的特征值提取方法能够提高电子邮件分类处理的正确率,并达到有效降低特征向量维数的目的。
关键词:电子邮件;文本分类;特征值提取
文章编号:1002-8331(2007)18-0177-04 **文献标识码**:A **中图分类号**:TP301

在这个互联网迅速发展的时代，电子邮件是一种有效、方便而又缩短路程的交流方式。它已经成为信息交互的重要工具,人们用它交流思想、传输文件、发表意见等。据 IDC 调查,2000 年全球日平均发送邮件超过 100 亿封,2005 年已经达 350 亿封以上。同时越来越多的商业公司和政府机构开始依靠电子邮件系统来从事商业和政务活动。

但是,不可避免的是用户定期会收到许多邮件,如何将大量的邮件归类整理成为日益严重的问题。特别是目前大量的垃圾邮件不断的充斥着电子邮件系统,这些垃圾邮件不仅给用户带来了不便,更为重要的是占用了大量网络资源。

尽管一些商业化产品在邮件到达用户端时,已采取一些措施过滤垃圾邮件。但传统的垃圾邮件过滤方法仅通过邮件的地址主题及手工输入的过滤条件来过滤邮件,显然随着邮件数目的不断增加，这种电子邮件过滤方法已不能满足人们的需求。国际数据公司调查表明,垃圾邮件被认为是因特网服务提供商(ISP)的第二大难题。

本文提出就是利用文本分类技术和全新的特征值提取方式来对电子邮件进行归类处理。这种方法不仅能有效地过滤出垃圾邮件,更为重要的是它能够将用户的邮件进行分类管理。

本文提出的特征值处理方法,特别针对邮件中的同义词问题作了特殊的处理。目前在文本分类中往往是把同义词和近义词当成一个词对待,这种方法处理起来非常方便和简单,但是对分类性能带来了很大的负面影响。而电子邮件由于它的随意性使得其中存在大量的同义词和近义词,如果仅将这些同义和近义词置之不理,那么将无法满足电子邮件分类系统高准确率的要求。

1 电子邮件的特性及特征值的初步提取

电子邮件除了具有文本文件的性质外,还有其特性[1]:

(1)Email 文本内容中往往包括大量的无关信息,例如敬意词,语气助词等;

(2)Email 中有非结构化数据,例如:文本、图片、声音等;

(3)Email 自身具有特定的结构,它包括发件人地址、收件人地址、主题、内容等;

(4)Email 附件的结构往往很难确定,例如 Word,Excel 等都很难从文件中直接读取数据来进行分析;

(5)Email 内容的不确定性,除了利用一定的规则进行通信,在对 Email 内容进行分析时很难通过一两个关键词就能确定 E-

基金项目:湖南省自科基金 (the Natural Science Foundation of Hunan Province of China under Grant No.02JJY2092);湘潭大学科研基金项目(No.06XZX02)。
作者简介:邹娟(1977-),女,讲师,研究方向:计算语言学;周经野(1948-),男,教授,研究方向:计算语言学;邓成(1978-),男,讲师,研究方向:模式识别;陈静(1968-),女,实验师。

mail 的类型。

(6)Email 长度的不确定性,由于电子邮件具有随意性,因此电子邮件的长度都是不确定的,但是一般来说,正文内容达到 2 000 字以上的大部分是些反动宣传或广告类型的垃圾邮件。

于是在本系统中首先针对电子邮件的特点,先对邮件的特征值进行了初步提取。

步骤 1 获取电子邮件的信息:将电子邮件的的主题、发件人、收件人、发件日期、邮件内容等从电子邮件中提取出来并保存到相应的数据库。去除图片、声音等非文本信息。

步骤 2 获取附件中的文本信息:将附件文档转化成标准的.xml 格式,然后再从 xml 文档中获取信息,保存到数据库中。

步骤 3 对电子邮件的地址进行过滤处理,目前普遍使用的电子邮件过滤系统多数采用的是地址过滤法,通过地址去除用户拒绝接受或经常发垃圾邮件的邮箱所发来的邮件。

步骤 4 建立了一个“停用词表”来过滤电子邮件中出现频率很高的敬意词、语气助词、虚词。然后利用词性特点将形容词以及一部分动词从文本中去除。

步骤 5 计算词 x_i 在每类邮件中的词频

$$p_{ij}=p(x_i|C_j)=\frac{x_i\text{ 在 }C_j\text{ 出现的次数}}{C_j\text{ 中特征词的总数}} \tag{1}$$

其中 C_j 表示类型,如 j=1 表示正常邮件类,j=2 表示非正常邮件类。得到词频后,我们删除了一些高频和低频的词,这些词往往对分类没有什么实际意义或意义不大。

步骤 6 不同位置词汇的加权处理。在人工阅读邮件时往往只要根据标题就能判断大部分的邮件的类型,所以出现在不同位置词汇对于分类的贡献是不一样的,于是引入了基数 B_j:

$$B_j=\frac{C_j\text{ 中特征词的个数}}{C_j\text{ 中的文本数}} \tag{2}$$

于是权值 $q_{ij}=k*B_j*p_{ij}$,其中 k 是词 x_i 在不同的位置取不同的值,如在实验中,如果 x_i 出现在标题中 k=0.08,如果 x_i 出现在第一段中 k=0.03,如果 x_i 出现在某段的第一句中 k=0.05,如果 x_i 出现在其它位置 k=0.01。

经过上面两步处理后就得到文本 C_j 的初始特征向量集 $\boldsymbol{X}=\{x_1,x_2,\cdots,x_{nj}\}$,和对应的权值向量 $\boldsymbol{Q}=\{q_1,q_2,\cdots,q_{nj}\}$。

2 电子邮件中的同义、多义词处理

在电子邮件当中,由于电子邮件的语言随意性使得对于同样的事物,在不同的情况下要用不同的词语来表达。于是出现了大量的同义词和近义词。因此,认为在不同类型的文本中(如垃圾邮件和正常邮件),同义词对于它们所属的共同概念的表达程度是不一样的,也就是说,具有一种模糊性。将这种模糊性用“模糊集合”来表达从而更加精确词汇之间的关系。

于是把一个同义(近义)词集合作为一个概念,定义为一个模糊集。利用文本分类是有导师学习的特点,在训练阶段利用分类中的训练样本自动计算每个词在该类型文本中与其对应概念的隶属度,形成一个模糊集。这样,在文本分类阶段,就可以利用前面得到的词汇在不同类型文本中对同义概念的隶属度来处理词汇。

下面具体地介绍同义词和多义词的处理方法。

2.1 基于模糊集的同义词处理

设若干个词汇 x^r,$1\leqslant r\leqslant m_j$,是同义词或者近义词,认为它们共同定义了一个概念。而这个概念可以用一个以这些词汇为论域的模糊集 S 来表示,即 $S=\{v^1/x^1,v^2/x^2,\cdots,v^{mj}/x^{mj}\}$,其中 v^r 是词汇 x^r 对于 S 的隶属度。

定义 1 设有若干个同义词或近义词 x^r,$1\leqslant r\leqslant m_j$,构成的同义词表 X 所形成的同义概念为模糊集 S,$S=\{(x^r,SRD(x^r,S))|x^r\in X,SRD(x^r,S)$为 x^r 隶属于 S 的隶属度$\}$。

在不同的文本类型中,词汇 x^r 对于同义概念 S 的隶属度是不同的。换言之,在不同的文本类型中同义概念 S 是不同的模糊集。所以,隶属函数 $SRD(x^r,S)$的值要因文本类型而定。

定义 2 设 C_i 为训练样本中的某类文本集,S_i 表示出现在 C_i 中的某个同义概念 S,而 x_i^r 为 S_i 中的某个词。定义 x_i^r 对于 S_i 的同义隶属函数为

$$SRD(x^r,S)=\frac{p_i^r}{\frac{1}{m_j}\sum_{r=1}^{m_j}p_i^r}\times g \tag{3}$$

其中 p_i^r 表示词频;比例系数 g,用来控制使 $SRD(x^r,S)\leqslant 1$(在实验中 g=0.1);词频

$$p_i^r=p(x_i^r|C_i)=\frac{x_i^r\text{ 在 }C_i\text{ 出现的次数}}{C_i\text{ 中特征词的总数}} \tag{4}$$

从上面的定义可以看出,$SRD(x^r,S)$实质上是 x^r 的词频与同属于 S 这个概念其它词的平均词频的一个比值。当然为了控制使 $SRD(x^r,S)\leqslant 1$ 引入了参数 g。

同义词处理一直是文本分类中的一个受到关注的问题,但是,由于自然语言词汇语义学方面的理论体系和技术还不完善,此问题一直没有得到很好的解决。本文的这种同义词处理方法一方面考虑了同义词和近义词之间的差别,针对不同类型的文本处理,使之更适用于文本分类词汇的处理。另一方面,系统中只用到了一般的同义词词典(实验中是根据梅家驹的“同义词词林”建立了一个小型的同义词词库)即可,而整个过程都由计算机自动完成,从而增强了系统的适应性和可扩展性。

2.2 基于模糊集的多义词处理

在对同义词的处理时,遇到另一个问题是,如果一个词具有多义时,那么它到底是和那个同义概念建立其同义关联呢?为了解决此问题,再次利用模糊集的知识处理。

设 $T^r=\{S^r(1),S^r(2),\cdots,S^r(n)\}$是多义词 x^r 的 f 个涵义组成的论域,其中 $S^r(k)$,$1\leqslant k\leqslant f$,是 2.1 节中的同义概念。则词 x^r 的多义性可以定义为论域 T^r 上的一个模糊集。同样考虑到训练文本集的区别,多义词 x^r 在 C_i 中的多义性为模糊集 $M_i^r=\{u_i^r(1)/S_i^r(1),u_i^r(2)/S_i^r(2),\cdots,u_i^r(f)/S_i^r(f)\}$,表示词 x_i^r 有语义特征 $S_i^r(k)$的隶属度为 $u_i^r(k)$。其中,$u_i^r(k)\in[0,1]$。

这里,主要就是对于训练样本中每类 C_i 中的词 x_i^r 对于对于每个词义的隶属函数。具体的计算方法如下:

$$u_i^r(k)=\frac{\sum_{\substack{e=1\\e\neq r}}^{m_i}p_i^e(k)}{\sum_{k=1}^{f}\left(\sum_{\substack{e=1\\e\neq r}}^{m_i}p_i^e(k)\right)} \tag{5}$$

其中,$p_i^e(k)$表示词 x_i^r 在 C_i 中与它第 k 个词义同义的第 e 个词的词频,即 $p_i^e(k)=p_i^e$。

从上面的公式,可以知道,$u_i^r(k)$实质上表现的是词 x_i^r 表现

不同语义(词义)的一个程度函数。

下面以实验中的实例加以说明。都知道 x_1^r="经济",它有两个意思:$S_1^r(1)$="节约:即用较少的人力、物力和时间获得较大的成果[9]";$S_1^r(2)$="本意,社会物质生产活动,或对国民经济有利或有害[9]"。

"经济"	
$S_1^r(1)$	$S_2^r(1)$
$p_1^1(1)=0.001\ 72\ \cdots\ p_1^4(1)=0.026\ 11$	$p_1^1(2)=0.029\ 61\ \cdots\ p_1^{14}(2)=0.091\ 27$
$u_1^r(1)=0.155\ 799$	$u_1^r(2)=0.844\ 201$

根据公式(3):

$$u_1^r(1)=\frac{(p_1^1(1)+\cdots+p_1^4(1))}{[(p_1^1(1)+\cdots+p_1^4(1))+(p_1^1(2)+\cdots+p_1^{14}(2))]}=\frac{(0.001\ 72+\cdots+0.026\ 11)}{[(0.001\ 72+\cdots+0.026\ 11)+(0.029\ 61+\cdots+0.091\ 27)]}=0.155\ 799$$

$u_1^r(2)$也是类似计算。

于是,对于"经济"来说在 C_1 中的多义性集 M_1^r={0.155 799/$S_1^r(1)$,0.844 201/$S_1^r(2)$}。

这样,对于多义词给出了对于每个词义的隶属度。于是,在分类中就可以根据词的不同词义加以计算。

这种多义词处理方法由于没有考虑过多的词汇语义学方面的内容,所以,使得处理起来比较容易,且易用计算机实现。

3　电子邮件分类系统设计模型

图 1 是邮件分类系统的结构图,对于未处理的邮件采用手工分类的方法来解决。

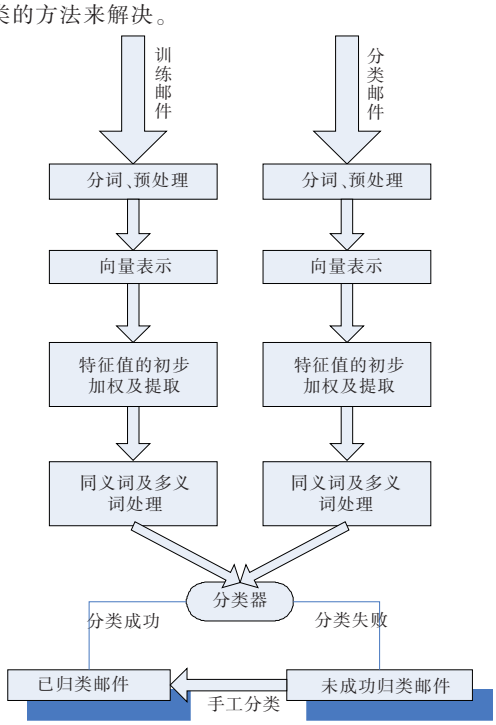

图 1　邮件分类系统处理模型

从图 1 中可以看出,在邮件分类系统中采用了分级处理的方法。在特征值的初步加权及提取阶段,采用的是第 1 章中提出的方法进行处理,得到文本的初始向量集合。到了同义词及多义词处理阶段采用了第 2 章提出的方法用模糊集来处理文本中的同义词和多义词,得到以"同义概念"为单位的特征值表示方法。

经过分类器分类后将未成功分类的邮件采用人工分类的方法,并把其结果反馈到特征值提取阶段,以指导后面的特征值提取。

4　比较实验

4.1　不同分类算法下的电子邮件分类比较实验

文本来源:电子邮件;

分词系统:ICTCLAS 系统;

分类类别:正常、垃圾邮件;

训练文本集:280 篇;

测试形式:开放式测试;

测试方法:(1)基于邮件的特征指提取方法的 SVM 分类系统;(2)基于邮件的特征指提取方法的 KNN 分类系统;

评价标准:正确率;

测试结果:如图 2 所示。

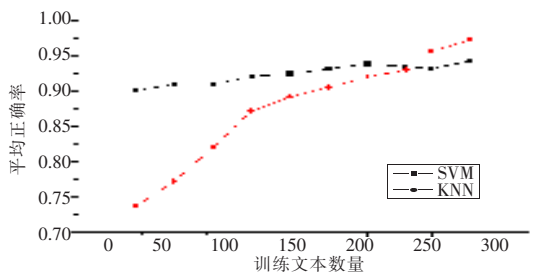

图 2　基于邮件的特征值提取方法在不同分类算法中比较图

测试分析:实验表明基于邮件的特征值提取方法在 SVM 和 KNN 两种分类方法中应用都达到了很好的分类效果。

4.2　同义概念为单位的特征值表示法与其它特征表示法在电子邮件分类中的比较实验

文本来源:电子邮件;

分词系统:ICTCLAS 系统;

分类类别:(1)正常邮件:经济、IT、体育;(2)垃圾邮件:广告、反动宣传;

训练文本集合:1 150 篇;

测试方法:(1)基于字;(2)基于词;(3)基于 2-grams;(4)基于同义概念;

测试评价指标:

(1)$F\text{-}measure=\frac{2\times准确率\times召回率}{准确率+召回率}$

(2)维数压缩率=$\frac{基于同义概念的特征值维数}{其它特征表示法的特征值维数}$

测试结果:如表 1 所示。

表 1　在电子邮件分类中基于同义概念的特征值表示与其它的特征值表示方法的比较　%

类型	名称	基于字		基于词		基于 2-grams		同义概念
		F-measure	压缩率	F-measure	压缩率	F-measure	压缩率	F-measure
正常邮件	经济	80.5		82.2		90.1		89.5
	IT	82.1	78.3	84.9	81.7	88.2	79.4	97.6 87.7
	体育	82.0	29.27	85.2	41.36	87.6	20.72	93.3
垃圾邮件	广告	85.7	91.6	87.1	85.4	86.3	85.4	99.3 97.5
	反动		86.2		88.2		84.7	99.1

测试分析：

(1)在测试中,不仅将电子邮件分为正常和垃圾邮件两类,还将其细分为小类,实验结果证明,即使对于小类特征值提取方法也能达到很好的分类效果。

(2)基于同义概念的特征表示方法在电子邮件分类中,其F-measure 明显要优于其它的特征表示法。

(3)基于同义概念的特征表示方法在电子邮件分类中,其特征值的数量明显要少于其它的特征表示法。

当然,由于分词的准确性,以及采取的只是用近似计算来分类,使得其分类的正确率还不是很好。这些方面将在今后的工作中加以改进。

5 结束语

本文针对电子邮件的特点提出了一种新的特征值提取方法,并将之有效地应用到电子邮件的分类当中,实验证明提出的特征值提取方法能够提高电子邮件的处理性能。接下来的工作将在文本分类算法上作进一步的研究工作,并将这种特征值表示和提取方法应用在其中。(收稿日期:2006 年 12 月)

参考文献：

[1] 朱炜,王晓国,黄韶坤,等.Email 挖掘系统的体系模型及其具体实现[J].计算机辅助工程,2004,2:1-10.

[2] Selamat A.Web page feature selection and classification using neural networks[J].Information Sciences,2004,158:69-88.

[3] Perrin P,Petry F E.Extraction and representation of contextual information for knowledge discovery in texts[J].Information Sciences,2003,151:125-132.

[4] 杨斌,孟志青.一种文本分类数据挖掘技术[J].湘潭大学自然科学学报,2001,23(4):34-37.

[5] 邹娟,周经野,邓成.一种基于语义分析的中文特征值提取方法[J].计算机工程与应用,2005,41(36):164-166.

[6] 谢宜辰.网络智能文本分类系统的研究与实现[J].湘潭大学自然科学学报,2000,22(1):12-15.

[7] Aseltine J H.Wave:an incremental algorithm for information extraction [C]//Proceedings of the AAAI,1998 Workshop on Machine Learning for Information Extraction,1999.

[8] 刘为国.Web 信息系统的体系结构[J].湘潭大学自然科学学报,2002,24(1):24-26.

[9] 中国社会科学学院语言研究所.词典编辑室现代汉语词典[M].北京:商务印书馆,2003.

[10] 朱红畅,孟志青.一种基于 SOM 和层次凝聚的中文文本聚类方法[J].湘潭大学自然科学学报,2005,27(4):36-39.

(上接 167 页)

[2] Oliveira S R M,Zaïane O R.Achieving privacy preservation when sharing data for clustering[C]//Proceedings of the International Workshop on Secure Data Management in a Connected World (SDM'04) in Conjunction with VLDB 2004,Toronto,Canada,August 2004.

[3] 张国荣,印鉴.应用等距变换处理聚类分析中的隐私保护[J].计算机应用研究,2006(7):83-86.

[4] Vaidya J,Clifton C.Privacy-preserving k-means clustering over vertically partitioned data [C]//Proc of the 9th ACM SIGKDD Intl Conf on Knowledge Discovery and Data Mining,Washington,DC,USA,August 2003:206-215.

[5] Merugu S,Ghosh J.Privacy-preserving distributed clustering using generative models [C]//Proc of the 3rd IEEE International Conference on Data Mining(ICDM'03),Melbourne,Florida,USA,November 2003:211-218.

[6] Jha S,Kruger L,McDaniel P.Privacy preserving clustering[C]//In 10th European Symposium on Research in Computer Security (ESORICS'05),Milan,Italy,September 2005:397-417.

[7] Clifton C,Kantarcioglu M,Vaidya J,et al.Tools for privacy preserving distributed data mining [C]//SIGKDD Explorations,2002,4(2):28-34.

[8] 罗永龙,徐致云,黄刘生.安全多方的统计分析问题及其应用[J].计算机工程与应用,2005,41(24):141-143.

[9] Du W,Zhan Z.Building decision tree classifier on private data[C]//Proceedings of the IEEE ICDM Workshop on Privacy,Security and Data Mining,Maebashi City,Japan,December 2002:1-8.

[10] Blake C L,Merz C J.UCI repository of machine learning databases[D].University of California,Irvine,Dept of Information and Computer Sciences,1998.

(上接 170 页)

[17] Miyauchi S.Collaborative multimedia analysis for detecting semantical events from broadcasted sports video[C]//16th International Conference on Pattern Recognition,Quebec,Canada,2002:1009-1012.

[18] So Zhong,Li Stan,Zhang Hong-Jiang.Extraction of feature subspaces for content-based retrieval using relevance feedback[C]//ACM Multimedia,Ottawa,Canada,September 30-October 5,2001.

[19] Naphade M R.A statistical modeling approach to content-based video retrieval [C]//IEEE Proceedings of 16th Internationgal Conference on Pattern Recognition,Quebec,Canada,2002:953-956.

[20] 王惠锋,孙正兴,王箭.语义图像检索研究进展[J].计算机研究与发展,2002(5):513-523.

[21] 余卫宇,余英林.视频语义信息的研究[J].计算机工程与应用,2004,40(6):27-29.

[22] 章毓晋.基于内容的视觉信息检索[M].北京:科学出版社,2003.

[23] 余卫宇,谢胜利,余英林,等.语义视频检索的现状和研究进展[J].计算机应用研究,2005(5):1-7.

[24] 蔡骏.基于语义的信息检索中的反馈技术[J].南京邮电学院学报,2003,23(2):78-81.

[25] 任和.语义视频对象的提取及其在视频检索中的应用[D].上海:复旦大学,2002.

[26] 张毅,赵捧未,刘怀亮,等.基于语义的图像相关反馈技术[J].情报杂志,2006,25(10):43-44.

[27] 庄越挺,潘云鹤,吴飞.网上多媒体信息分析与检索[M].北京:清华大学出版社,2002.

[28] 张若英,申铉京.基于内容的视频检索方法的研究[J].计算机工程与应用,2004,40(6):196-199.