DOI:10.13203/j.whugis2004.12.019
第29卷 第12期 武汉大学学报·信息科学版 Vol.29 No.12
2004年12月 Geomatics and Information Science of Wuhan University Dec. 2004

文章编号:1671-8860(2004)12-1115-04 文献标识码:A

基于智能决策支持系统的电子邮件过滤技术

张沪寅[1] 张文熙[1] 吴产乐[1] 邢建兵[1]

(1 武汉大学计算机学院,武汉市珞喻路129号,430079)

摘 要:通过分析实现邮件过滤技术,提出了基于智能决策支持系统的电子邮件过滤模型,并重点讨论了在模型中如何通过内容分析实现对邮件的分类。

关键词:智能决策支持系统;垃圾邮件;过滤模型;质心分类算法;邮件分类

中图法分类号:TP393

过滤技术是目前反垃圾邮件用到的主要技术。通过对邮件内容的分析实现电子邮件过滤一般常用关键词匹配的方法,对邮件进行检索,根据关键词出现的频率来判断一个邮件是否为垃圾邮件,以便进一步处理。本文通过对邮件的内容分析,提出了基于智能决策支持系统的电子邮件过滤模型。由于实现针对内容进行过滤的关键是邮件的分类,因此,在模型中重点讨论了如何通过内容分析实现对邮件的分类。

1 基于IDSS的电子邮件过滤模型设计

基于IDSS(intelligent decision support systems)的电子邮件过滤模型由电子邮件接收模块、过滤模块、垃圾邮件数据库、模型库及其管理模块、知识库及其管理模块、用户接口共8个部分组成,其中模型库及其管理模块、知识库及其管理模块分别对应IDSS中的模型库子系统和知识库子系统。

1.1 电子邮件接收模块

电子邮件接收模块主要实现对电子邮件的实时接收,它不断地检测邮件接收端口(如对于邮件接收服务器,其端口号通常为25)。当邮件接收端口接收到电子邮件后,邮件接收模块从邮件接收端口获得电子邮件,然后把该邮件保存到电子邮件接收队列的队尾。当电子邮件接收队列中有邮件时,把队首的邮件转发到过滤模块。

1.2 过滤模块

过滤模块包括邮件解析、内容分析、邮件分类3个子模块。

邮件解析子模块。电子邮件在传输过程中都已被编码,接收到的邮件需要被解码才能被用户识别,电子邮件的编码方式参见文献[1,2]。邮件解析子模块的作用就是把电子邮件解码,使其成为内容分析子模块能够识别的形式。模块根据电子邮件的信息格式,识别出邮件文本的编码类型,并从邮件及其附件中提取出正文的文本内容。例如,8bit或7bit的ASCⅡ编码的文本不需要解码。对于quoted-printable编码和base64编码的内容,根据编码方法将其解码。

内容分析子模块.内容分析子模块通过分析邮件的内容,为邮件的分类做准备。对已解码的文本进行关键词的检索/扫描,找出所有能反映邮件内容的关键词,并按照组成特征向量必需的所有分量,把邮件表示成分类算法需要的向量形式。

邮件分类子模块。邮件分类子模块依据模型库中的分类规则,把邮件分类成垃圾邮件或合法邮件。邮件分类子模块调用模型库中的分类算法对邮件的特征向量分析计算,与知识库提供的所有垃圾邮件的特征向量作比较,计算出它们之间的相似度,然后把相似程度符合分类规则的邮件划分成垃圾邮件。

在邮件分类子模块把电子邮件分类后,过滤模块将合法的邮件送往邮箱,同时将垃圾邮件保存到垃圾邮件库。对于相似度接近分类规则而难以判断的邮件,可以将邮件信息发送给用户接口,

收稿日期:2004-10-10。

项目来源:国家863计划资助项目(2003AA001032);湖北省自然科学基金资助项目(2001B057)。

并给出建议,由用户确定是否为垃圾邮件。

1.3 垃圾邮件数据库

垃圾邮件数据库是储存垃圾邮件的数据库。在邮件过滤的实现中,可能出现将合法邮件误判为垃圾邮件的情况,而且当一个新的垃圾邮件出现后,还需要对其特征进行分析,通过新垃圾邮件的特征对知识库进行更新。所以过滤后的垃圾邮件不能立即删除,而应该储存在垃圾邮件库中。

1.4 知识库及其管理模块

知识库储存垃圾邮件特征向量集,也可称为垃圾邮件特征向量库。垃圾邮件特征向量集是代表垃圾邮件特征的特征向量的集合,它为邮件的比较提供依据。垃圾邮件都有一些共同的特征,例如频繁出现的某些关键词等。收集一批垃圾邮件作为垃圾邮件样本,把这些垃圾邮件样本的共同特征提炼出来,作为组成特征向量所需要的所有分量,把所有的垃圾邮件样本表示成这种特征向量的形式,然后储存在知识库中。知识库中的垃圾邮件特征向量集是过滤模块中用于比较的对象,同时也是模型库中分类规则制定的数据源。

知识库管理模块的输入数据是垃圾邮件数据库中的垃圾邮件,模块的主要作用是建立和更新知识库。首先,收集一批有代表性的垃圾邮件,管理模块用其特征向量形成初始的知识库。当一个新的垃圾邮件出现后,管理模块依据其特征更新知识库。知识库管理模块可以自动地建立或更新知识库,也可以通过用户接口人工建立或更新知识库。

1.5 模型库及其管理模块

模型库包括实现邮件分类的分类模型,即分类算法,以及分类规则。

模型库管理模块的作用是建立和维护模型库,它一般由用户通过用户接口人工操作,分类规则也可以由该模块自动修正。

1.6 用户接口

用户接口是用户与邮件过滤系统交流的图形界面,功能主要有:建立和查询所有的数据库;管理和维护知识库和模型库;发现垃圾邮件数据库中的合法邮件;处理过滤模块的反馈信息,对过滤模块难以判断的邮件给出建议显示给用户;用户通过用户接口对过滤模块难以判断的邮件作出判断;为用户提供帮助信息等。

2 模型库子系统

模型库子系统包括模型库和模型库管理模块,它为过滤模块提供实现邮件分类的分类算法和分类规则。

2.1 分类算法

目前,用于电子邮件过滤的分类算法主要有 K 邻域方法(K-nearest neighbor classification)[3]、朴素贝叶斯方法(naive Bayesian classification)[4] 等,本文介绍一种基于质心的邮件分类算法——质心分类算法(centroid-based classification)[5]。

2.1.1 质心分类算法

质心分类算法的基本思想是:用同一类电子邮件的质心代表其共同特征,需要被分类的邮件与质心比较,如果该邮件与质心的相似程度满足分类要求,则该邮件就可以被划分成质心所代表的这一类电子邮件。

在质心分类算法中,数据项由空间向量模型表示,邮件被表示为一个字或词出现频率的向量:$e_f=(f_1, f_2, \cdots, f_n)$,其中 f_i 是第 i 个词在邮件中出现的频率;n 是代表一类邮件关键词的数量。由于一些词出现在许多邮件中,为了突出这些词对不同邮件的代表作用,通常对不同邮件的 f_i 都乘以一个权值 $\log(N/\mathrm{d}f_i)$,其中 N 是集合中邮件的总数,$\mathrm{d}f_i$ 是包含第 i 个词的邮件数。于是,邮件被表示为:$e_f=(f_1\log(N/\mathrm{d}f_1), f_2\log(N/\mathrm{d}f_2), K, f_n\log(N/\mathrm{d}f_n))$。然后对不同长度的邮件进行标准化,使它们成为单位向量,即 $\| e_f \|_2=1$。

在空间向量模型中,两个邮件 e_i 和 e_j 的相似度用两个邮件向量的夹角的余弦表示:

$$\cos(e_i, e_j)=\frac{e_i \cdot e_j}{\| e_i \|_2 \times \| e_j \|_2} \tag{1}$$

因为邮件向量是单位向量,所以式(1)可表示为 e_i 和 e_j 向量积:$\cos(e_i, e_j)=e_i \cdot e_j$。

用 E 作为一个邮件集合$\{e\}$的向量表示,e 作为邮件的向量表示,则质心向量 C 定义如下:

$$C=\frac{1}{|E|}\sum_{e\in E} e \tag{2}$$

质心向量 C 可以看作是具有某种特征的一组向量的平均值,在邮件分类中,它用来表示邮件集合$\{e\}$的共同特征。

与邮件之间的相似度类似,邮件向量 e_f 和质心向量 C 的相似度也可以用其夹角的余弦表示:

$$\cos(e_f, C)=\frac{e_f \cdot C}{\| e_f \|_2 \times \| C \|_2}=\frac{e_f \cdot C}{\| C \|_2} \tag{3}$$

虽然邮件向量是单位向量,但是质心向量不一定是单位向量,即 $\| C \|_2$ 不一定等于 1。

式中,

$$e_f \cdot C = e_f \cdot \left(\frac{1}{|E|} \sum_{e \in E} e \right) = \frac{1}{|E|} \sum_{e \in E} e_f \cdot e = \frac{1}{|E|} \sum_{e \in E} \cos(e_f, e) \quad (4)$$

$$\| C \|_2 = \sqrt{C \cdot C} = \sqrt{\frac{1}{|E|^2} \sum_{e_i \in E} \sum_{e_j \in E} \cos(e_i, e_j)} \quad (5)$$

2.1.2　与其他分类算法的比较

质心分类算法用质心向量总结了每一类邮件的普遍特征, 质心向量中一些突出的维(例如权重高的词)只需要反映在这些词出现频繁的邮件中, 而不需要在同一类所有邮件中反映, 这一点对于高维集合尤其重要。对邮件特征的总结, 质心分类算法(包括朴素贝叶斯法)要比 K 邻域方法执行得更好。

在计算单个邮件与一类邮件的相似程度上, 质心分类算法比朴素贝叶斯算法有更好的性能。朴素贝叶斯算法中假定不同的词在邮件中的出现是相互独立的, 这种假定在实际的邮件中是不成立的, 词之间的相互依赖将导致朴素贝叶斯算法作出不恰当的估计, 使得判断邮件是否属于某类邮件时出现错误。

质心分类算法使用简单的函数说明词之间的依赖性, 新邮件和某类邮件的相似度通过两项(式(4)和式(5))的比值计算, 前一项和朴素贝叶斯算法中的可能性估计非常相似, 但第二项却说明了词之间的依赖性。通常, 如果第二项的值较高, 就说明邮件中词的相互依赖出现的程度较高; 如果第二项的值低, 就说明邮件中词的相互依赖出现的程度也低。第二项实际上扮演了一个对相似度过高(或过低)估计的纠正参数。

2.2　邮件分类

实现垃圾邮件过滤的邮件分类就是从邮件中搜索出能代表邮件内容的一些关键字或词[6], 使用分类算法(本文使用质心分类算法)来计算邮件和垃圾邮件的相似度, 然后依据分类规则确定邮件是否是垃圾邮件。

2.2.1　邮件的分类方法

利用质心分类算法对邮件进行分类的过程如图1 所示。

邮件的分类过程是一个不断循环的过程, 设垃圾邮件特征向量集为$\{ e_i, i = 1, 2, \cdots, n \}$, 则邮件分类的步骤如下。

1) 读取解码后邮件 e 的文本内容;

2) 从邮件的文本内容中搜索出代表该邮件内容的关键词, 计算出每个关键词的出现频率 f_i, 由垃圾邮件特征向量集提供的数据计算与每个关键词对应的权值 $\log(N/\mathrm{d}f)$, 然后按照预先制订的邮件向量所需的关键词对邮件向量化, 得到邮件的向量表示 $e_f = (f_1 \log(N/\mathrm{d}f_1), f_2 \log(N/\mathrm{d}f_2, K, f_n \log(N/\mathrm{d}f_n))$, 其中在邮件 e 中没有出现的关键词所对应的分量用 0 代替;

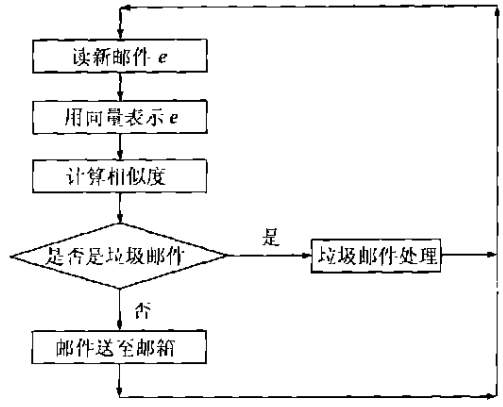

图 1　分类流程图

Fig. 1　Classification Flow

3) 计算邮件向量 e_f 和垃圾邮件质心向量 C 的相似度 $\cos(e_f, C)$;

4) 用相似度 $\cos(e_f, C)$ 与临界值 K 作比较(临界值 K 在制定分类规则时给出), 把满足条件 $\cos(e_f, C) \geqslant K$ 的邮件划分为垃圾邮件, 不满足该条件的邮件划分成合法邮件;

5) 若邮件 e 被划分成垃圾邮件, 则将邮件 e 暂时保存供管理员处理, 并且用邮件 e 作为新的垃圾邮件标本更新垃圾邮件特征库, 然后转到步骤 7);

6) 若邮件 e 被划分成合法邮件, 则将邮件 e 送至邮箱;

7) 等待新的邮件出现, 当新的邮件出现时, 重复步骤 1) ~ 步骤 6)。

2.2.2　分类规则的制定

分类算法只能确定邮件与垃圾邮件的相似程度(或与合法邮件的相似程度), 是否把邮件划分成垃圾邮件需要分类规则的指导。分类规则的制定首先要确定临界值 K, 临界值 K 的确定需要用到垃圾邮件特征向量集。设垃圾邮件特征向量集为$\{ e_i, i = 1, 2, \cdots, n \}$, n 是垃圾邮件特征向量集中垃圾邮件特征向量的个数, e_i 为垃圾邮件特征向量, 以下是临界值 K 的确定方法。

1) 计算垃圾邮件质心向量 C;

2) 计算每个垃圾邮件特征向量与垃圾邮件

质心向量 C 的相似度 $\cos(e_i, C)$;

3) 用 $K = \frac{1}{n}\sum_{i=1}^{n}\cos(e_i, C)$ 计算临界值 K。

制定的分类规则要使得合法邮件被划分成垃圾邮件的概率尽可能小, 其分类规则为:满足条件 $\cos(e_i, C) > K$ 的邮件划分为垃圾邮件, 其中 e_f 是需要被分类的电子邮件的向量表示。在经过对一些电子邮件分类后, 垃圾邮件特征库中又会积累一批新的垃圾邮件特征向量, 可以利用这些新增的向量重新计算临界值 K, 该功能通常由规则库管理模块实现。

3 模型的特点分析

1) 由于模型是基于智能决策支持系统的, 所以用户(或管理员)也成为整个系统的一部分。作为决策支持系统的一部分, 用户能够参与到对电子邮件的过滤, 对于难以自动过滤的邮件, 用户可以根据系统提供的建议方案, 用人工处理, 这样可以减少过滤的错误率。

2) 质心分类算法充分考虑了关键词之间的依赖性, 用该算法作为分类算法可以在很大程度上减少分类错误的可能性。

3) 用垃圾邮件数据库保存垃圾邮件, 能把被误判为垃圾邮件的合法邮件人工发送到邮箱, 保证合法邮件不被错误删除;同时也避免了用合法邮件去更新垃圾邮件特征库。

4) 模型中制定的分类规则既能使合法邮件被划分成垃圾邮件的概率尽可能小, 又能够保证过滤效率不至于太低。

参 考 文 献

1 谢希仁. 计算机网络(第三版). 大连:大连理工大学出版社, 2000

2 Tanenbaum A S. 计算机网络(第三版). 熊贵喜, 王小虎, 译. 北京:清华大学出版社, 1998

3 Yang Y, Liu X. A Re-examination of Text Categorization Methods. SIGIR-99, University of California at Berkeley, 1999

4 McCallum A, Nigam K. A Comparison of Event Models for Naive Bayes Text Classification. AAAI-98 Workshop on Learning for Text Categorization, Madison, Wisconsin, 1998

5 Han E H, Karypis G. Centroid-based Document Classification Algorithms: Analysis & Experimental Results. Technical Report TR-00-017, Department of Computer Science, University of Minnesota, Minneapolis, http://www.cs.umn.edu/~karypis, 2000

6 Nuanwan S, Kanokwan C, Piyanan T. Anti-Spam Filtering: A Centroid-Based Classification Approach. ICSP'02, Beijing, 2002

第一作者简介:张沪寅, 副教授, 博士生, 现从事计算机网络安全、网络管理和现代数据库的研究。

IDSS-based E-mail Filtering

ZHANG Huyin[1] *ZHANG Wenxi*[1] *WU Chanle*[1] *XING Jianbing*[1]

(1 School of Computer Science, Wuhan University, 129 Luoyu Road, Wuhan 430079, China)

Abstract: While it brings us large convenience, E-mail results in a new problem, junk mails. Mail-filtering is filtering junk mails from many E-mails. This paper introduces E-mail filtering by analyzing the content of mail and gives the model of based-IDSS E-mail filtering. This paper also discusses how to classify E-mails by analyzing the content of mails in the model.

Key words: IDSS; spam; filtering model; centroid-based classification; mail

About the first author: ZHANG Huyin, associate professor, Ph.D candidate, majors in computer network security, network management and modern time database.

(责任编辑: 晓晨)