基于 IDSS 的中文垃圾邮件过滤模型设计

龚 伟,李柳柏
(涪陵师范学院 计算机科学系,重庆 408003)

摘 要:以智能决策支持系统结构为基础,提出了一种新的电子邮件过滤模型,并对中文垃圾邮件过滤中的中文分词及垃圾邮件特征知识库的更新等关键问题进行了探讨。开发了"智能邮件过滤系统(IEFS)",使垃圾邮件误判率得到了一定程度的控制,有效防止了垃圾邮件的泛滥。

关键词:IDSS;垃圾邮件;知识库;分词
中图分类号:TP393.098 **文献标识码**:A **文章编号**:1673-629X(2007)03-0163-03

Chinese Spam Mail Filtering Model Design Based on IDSS

GONG Wei, LI Liu-bai
(Department of Computer Science, Fuling Normal University, Chongqing 408003, China)

Abstract: Based on structure of intelligent decision support system, it put forward a kind of new E-mail filtering model, and discussed the key problem such as Chinese word segmentation and spam characteristic knowledge base updating, etc. And have developed "Intelligent E-mail Filter System (IEFS)". It has made the rate of spam erroneous judgment under the control, efficiently avoid spam overflow.

Key words: intelligent decision support system; spam; knowledge base; word segmentation

0 引 言

伴随着互联网技术的不断发展,垃圾邮件日益泛滥,根据 IDC 最新的调查统计结果显示,2005 年,垃圾邮件的增长速度是 E-Mail 用户增长速度的 3.5 倍左右。过多的垃圾邮件占用了大量的网络带宽和空间资源,严重时将会拥塞整个 Internet 链路,中断 Internet 的部分线路的运营而造成巨大的经济损失,更为严重的是,如果用户不小心打开了携带病毒的垃圾邮件,其危害可能是致命的。

反垃圾邮件的战争已成为一场全面的持久战,吸引了大量的研究人员从各个方面进行深入研究,而反垃圾邮件技术的核心仍在于垃圾邮件过滤方法的研究。目前使用较多的垃圾邮件自动过滤方法可以归纳为两大类:一类是基于规则的邮件过滤,主要是利用包含各种约束条件的规则集来过滤,如决策树(Decision Tree)、粗糙集(Rough Set)等方法;另一类是基于统计的邮件过滤,包括贝叶斯(Bayes)、K-最近邻(K-Nearest Neighbor, K-NN)算法以及支持向量机(Support Vector Machines, SVM)等方法。但无论是基于规则还是基于统计的方法,其基础都是针对邮件的内容和特征,利用关键词匹配的方法,对邮件进行检索。垃圾邮件制造者也挖空心思,在一些中文敏感词上做文章,采用在敏感词中间插入其它符号或将某个字替换为同音字等手段,来逃避邮件过滤器的"搜捕"。

收稿日期:2006-05-21
基金项目:国家教育部"春晖"计划(Z2005-1-55003)
作者简介:龚 伟(1977-),男,重庆人,讲师,硕士,研究方向为计算机网络安全及决策支持系统;李柳柏,副教授,研究方向为计算机网络安全。

1 基于 IDSS 的邮件过滤模型设计

1.1 IDSS 结构

IDSS[1](Intelligent Decision Support System,智能决策支持系统)是在决策支持系统(DSS)基础上集成人工智能的专家系统(Expert System, ES)而形成的。决策支持系统主要由人机交互与问题处理系统、模型库系统(由模型库管理系统和模型库组成)、数据库系统(由数据库管理系统和数据库等组成)组成。专家系统主要由知识库、知识库管理系统和推理机三者组成。决策支持系统和专家系统集成为智能决策支持系统,其具体集成结构形式如图 1 所示。

1.2 基于 IDSS 的邮件过滤模型设计

以 IDSS 结构为基础,笔者设计了一种新的电子邮件过滤模型(如图 2 所示),新的邮件过滤模型主要由邮件过滤模块、用户接口、邮件分类模型库及 MBMS

(模型库管理系统)、垃圾邮件数据库及 DBMS(数据库管理系统)、垃圾邮件特征知识库及 KBMS(知识库管理系统)五大部分组成。

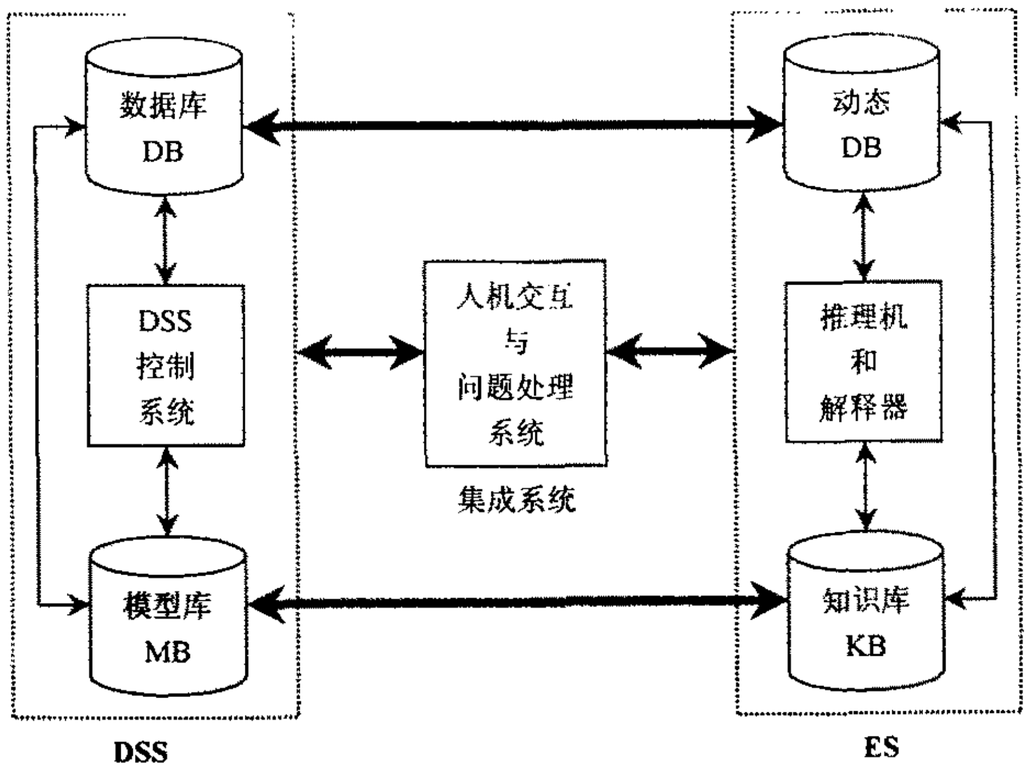

图 1 智能决策支持系统集成结构图

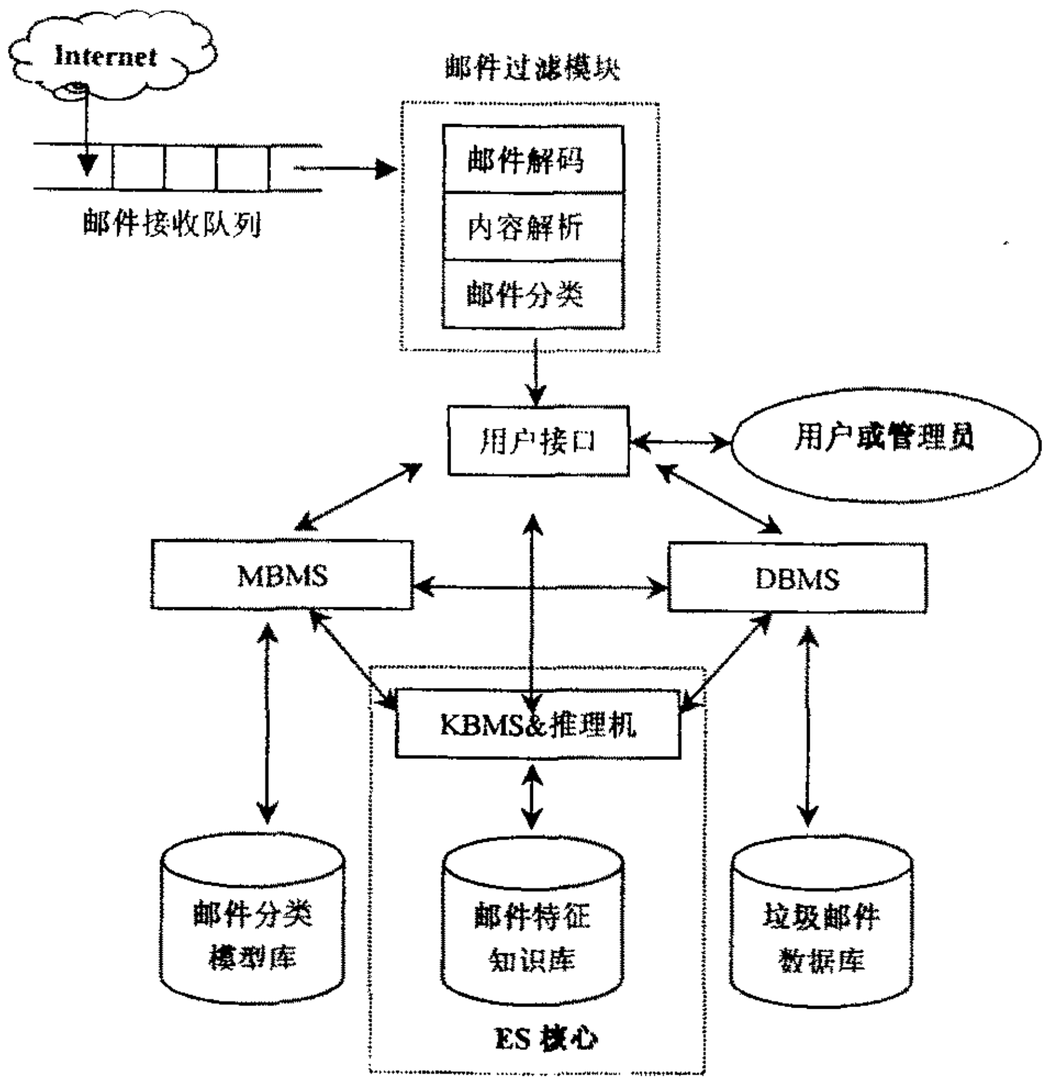

图 2 基于 IDSS 的邮件过滤模型

邮件接收程序不断检测邮件接收端口,将获得的电子邮件保存到邮件接收队列的队尾,当邮件接收队列中有邮件时,将队首邮件转发到邮件过滤模块。邮件过滤模块首先根据邮件的信息格式,识别出邮件文本的编码类型,然后对接收的邮件进行解码,解码后的文本直接进行内容分析,如检索出所有能反映邮件内容的关键词,以供邮件分类子模块调用邮件分类模型库中的分类规则和分类算法进行分析计算,过滤出垃圾邮件。经过邮件过滤模块分类后,邮件被分成了三类:合法的邮件,通过用户接口直接送到用户邮箱。垃圾邮件,送往垃圾邮件数据库,通过对其特征进行分析,更新垃圾邮件特征知识库,垃圾特征知识库为以后的内容检索、比较、分析提供依据。KBMS 可以通过提供的垃圾邮件训练样本集来初始化和更新垃圾邮件特征知识库,也可以通过用户接口人工建立和更新。还有一类是垃圾邮件特征不明显、根据分类规则难以判断的邮件,可以通过用户接口,提交给用户进行判断,如用户认为是垃圾邮件,仍要将其送入垃圾邮件数据库,提取特征向量,更新垃圾邮件特征知识库。

由于模型是基于智能决策支持系统的,所以用户(或管理员)成为了系统的一部分,对于难以自动过滤的邮件,系统会通过用户接口提供建议给用户,由用户决策。这可以减少过滤的错误率。同时,垃圾邮件数据库的存在为被误判为垃圾邮件的合法邮件的人工恢复提供了可能,也可以避免用合法邮件去更新垃圾邮件特征知识库。

2 邮件过滤中特殊问题的处理

2.1 中文邮件过滤的难点

邮件过滤模块中的内容分析子模块通过分析邮件内容,为邮件分类做准备。对已解码的中文文本进行关键词的检索,找出所有能反映邮件的关键词。对英文关键词的检索过程比较容易,方法较多,因为英文邮件每个单词之间有空格,即使垃圾邮件制造者对英文关键词使用过去分词、复数形式或将词用一些符号分割为几段,也可很容易地根据词态学方法对邮件内容进行还原,但基于内容的中文邮件过滤难度较大,特别是中文自动分词问题,一直制约着中文文本检索技术的发展,这是由于在中文句子中,词与词之间没有任何分隔符,而且中文的词法约束很不规范,千变万化,加之中文本身的二义性,更是增加了中文分词的难度。

另一方面,垃圾邮件制造者不断更新反过滤方法,给邮件过滤带来了更多困难,比较典型的做法是在某些敏感词中间插入不等个数的无意义字符,或者使用中文同音字代替敏感词,以对抗过滤措施[2,3]。

2.2 中文分词法

使用一种规则与统计相结合的分词方法来完成中文分词。这个方法在基本分词方法基础上,既不完全依赖于语法语义规则,也不是完全的语料库统计,其主要思想是:调用正向最大匹配法和逆向最大匹配法对中文句子进行切分,如果两者一致就是正确的,否则通过比较词的个数、未登录词以及词频来选择正确的切分,若仍然不能解决问题,则根据规则确定歧义字段的切分,这里的规则是从一个标注了的语料库(http://icl.pku.edu.cn /icl-groups/)采用统计方法得到。具体分词步骤如下:

(1)改进的正向和逆向最大匹配法分词[4]。

传统的正向和逆向最大匹配法分词是分别从整个

字串两端开始按长词优先原则进行匹配,每次去掉可匹配的最长词,再对剩下的字符串重复上述过程直至分词结束。但由于汉语中双音节词占70%以上,所以对最大匹配算法进行了改进:分词时先读入两字,然后看词典中有没有包含这两个字的词,有就增加一个字,直到增加后词典中不存在为止,这时所取得的词就是最大匹配得到的词,若开始就查找不到包含初始两个字的词,那两个字的首字就作为一个词(为未登录词)。对于正向与逆向都是这种算法,并且该步骤结果中还要带上各个词所有可能的词性及其对应的词频(Word Frequency,WF),未登录词须注明。若结果一致,则结果就作为正确的分词结果;否则,则找出不一致的最小字段,对每一个不一致的字段继续下面的步骤。

(2)词数—未登录词—词频比较。

比较由正向最大匹配法分词和逆向最大匹配法分词后的词数,取词数较少的一个作为结果,因为据统计组成长词的可能性比例很高;若词数相等,则比较未登录词数,以未登录词的个数少的一个作为结果;若未登录词的个数相等,则先取出两个字段的每一个词(未登录词除外)的词频,比较它们中的词频最小的词,若差值大于或等于某一阈值,则取词频最小的词较大的那一个字段作为结果。如:结果"A/BC"与结果"AB/C",且有词频 WF(A) <F(BC),WF(AB)>WF(C),WF(A)-WF(C)≥阈值 CW,则选择"A/BC"作为结果。若小于阈值,则比较两个字段的词频总和(未登录词词频记为0):若词频总和之差大于或等于某一阈值,则取词频之和大的作为正确的分词结果。如上面的两个结果中:词频 WF(A)<WF(BC),WF(AB)>WF(C),但是 WF(A)-WF(C)<阈值 CW,就转为比较 WF(A)+WF(BC)跟 WF(AB)+WF(C),若(WF(A)+WF(BC))-(WF(AB)+WF(C))≥阈值 CS,则选择前者即"A/BC"作为结果。若结果小于某一阈值,则转至下一步。

(3)查找规则库。

规则库记录的是词与词之间的邻接概率,分为两类:个性规则和共性规则。个性规则表示具体词之间的邻接可能性,共性规则给出词类之间的邻接概率。个性规则分成两种形式:

①"词/词",如规则"要/强调"把分词结果"要强/调"排除掉了。

②"词/POS"或"POS/词",POS表示词性。某个具体的词可与某个词性的词前接或后接。共性规则的形式为 POS1/POS2,POS1 为前一个词的词性,POS2 为后一个词的词性。使用规则库时,先查找个性规则,再查共性规则。如果以上方法都不能解决问题,就采用人工干预,同时将这个特殊的取舍结果存到规则库中作为个性规则,供下次使用。

2.3　垃圾邮件特征关键词的检索

垃圾邮件中往往包含一些特定的诸如"赚钱"、"免费"之类的关键词,垃圾邮件制造者将一些敏感词中插入无意义字符后,导致中文分词方法失灵。针对这一问题,一种做法是将中文文本中的所有无意义字符去掉,再进行中文分词,但去掉所有无意义字符将意味着文本原有的赖以断句的标点符号也将被去掉,整个文本成为一个大"句子",加大了中文分词的难度;另一种可行的方法是首先通过学习建立一个垃圾邮件特征词词库(属于垃圾邮件特征知识库的一部分),在针对中文邮件文本提取词时,一旦发现某个字是垃圾邮件特征词词库中的某个词语的首字,即在随后6个字范围内检索该词语的第二个匹配的字,如找到,再对下文进行类似的检索,直到完成整个词的匹配,完成匹配后,即对敏感词进行还原(删除插入的无意义字符),以此作为中文分词的预处理[5]。

2.4　垃圾邮件特征知识库的更新

笔者收集了2788封垃圾邮件作为垃圾邮件训练样本集供系统学习,学习过程中,如果使用2.3节中的方法匹配出了某个敏感关键词,将这个词的词频扩大2倍;如果某个词出现在邮件的 Subject 中,词频扩大3倍。统计出所有词的词频后,由高到低排序,找出词频最高的3~5个词或短语,作为该垃圾邮件的主题词,添加到垃圾邮件特征知识库中。

3　实验结论

根据图2的模型设计,完成了"智能邮件过滤系统(IEFS)"的开发,从集成结构形式上看,属于一种 DSS 为主体的 IDSS 结构,以定量分析为主体,结合定性分析来解决垃圾邮件过滤问题。智能邮件过滤系统目前已在本校邮件服务器上试运行,从运行效果来看,邮件误判率得到了一定程度的控制,部分原因是由于用户决策的引入,同时也得益于垃圾邮件特征知识库的不断更新。但在试用了一段时间后,发现垃圾邮件的漏检率有小幅上升,一方面是由于抵抗反垃圾邮件手段不断变化,对关键字检索过滤的方法也需要推陈出新,不断改进;另一方面,不同的用户对垃圾邮件的界定也不完全相同,例如某个企业对发到其邮箱里的与其产品相关的广告认为是合法邮件,而对一般用户而言,只要是没经过订阅的广告均被认为是垃圾邮件。

IEFS 有效控制了校园网内部垃圾邮件的泛滥,笔者将在对其知识库进行完善的基础之上,逐步推向市场,使其能在更大范围内发挥作用。　(下转第241页)

路调整为从通辽支点通过,在此基础上再次求解最短路径。由于一条线路可能满足多条经由要求,故需将新得到的调整路线的所经站点与经由要求文件对比扫描,如此反复直到没有经由要求满足此线路为止。

算法描述如下:

(1)输入起始站和终点站以及列车的品类名;

(2)对发到站和列车的品类名进行合法性检查,并查询发到站对应的分局号和列车品类号;

(3)求解发到站之间的最短路径;

(4)结合发到站对应的分局号和列车品类号进行经由要求扫描,检查是否有符合条件的经由要求,若存在则将改走线路写入线路调整表;

(5)对原最短路径的途径各站点进行经由要求扫描,检查是否有符合条件的经由要求,若存在则将改走线路写入线路调整表;

(6)根据线路调整表中的数据重新确定发到站,返回执行第 3 步,直到线路调整表无新数据添加。

3　特定径路计算实例

计算径路(见表 1):哈尔滨 - 成都, 一般重车

经过将原路线反复与经由要求对比扫描,可以得到以下几条经由要求:

①特定经由:凡经山海关支点装到成都局成都的重车,经京秦线,按丰台支点运输。

②特定经由:凡经丰台支点装到成都局成都分局的重车,经京广、新焦、焦柳线,按襄樊北支点运输。

③特定经由:凡经安康支点装到成都局成都分局的重车,经阳安线、宝成线广元分界站运输。

首先计算最短径路:哈尔滨 - 成都

表 1　原径路节点里程表

站名	哈尔滨	长春	沈阳	锦州	山海关	秦皇岛	丰台
里程(km)	0	242	547	783	967	983	1296
站名	石家庄	侯马	新丰镇	咸阳	宝鸡	阳平关	成都
里程(km)	1555	2081	2361	2420	2572	2843	3241

扫描特定经由定义:

特定经由:京秦线:山海关 - 丰台

特定经由:京广、新焦、焦柳线:丰台 - 新乡 - 焦作北 - 襄樊北

特定经由:阳安,宝成线:安康 - 广元

经调整计算最终车流径路(见表 2):哈尔滨 - 成都

表 2　径路调整后节点里程表

站名	哈尔滨	长春	沈阳	锦州	山海关	秦皇岛	丰台
里程(km)	0	242	547	783	967	983	1296
站名	新乡	焦作北	襄樊北	老河口	安康	阳平关	广元	成都
里程(km)	1894	1956	2463	2511	2822	3179	3358	3577

这一算法定义和计算车流径路的方法不仅直观、清晰、易用, 而且易于检查与维护。

4　应用前景

车流径路是编制货物列车编组计划最主要的依据之一,在铁道部的各类处理系统如货票制票、资金清算、精密客货统计、成本核算、营销支持、收入审核检查中都不同程度地存在经由计算,因此,车流径路方案的选择和车流径路管理一直是铁路运营管理工作倍受关注的问题之一,它是编制货物运输计划、技术计划、列车编组计划以及列车运行图的基础,同时也是进行车流推算、清算各铁路局货运收入、对发货人核收运费的依据。研究铁路运输特定经路算法,实现了最短路径及特定径路的求解, 提高了工作效率,保证了正确性。

参考文献:

[1] Dijkstra E W. A note on two problems in connexion with graphs[J]. Numerische Mathematik, 1959(1):269 - 271.

[2] 施昌明. 全国站间最短径路特定经由里程算法的电脑实现[J]. 电脑开发与应用, 1993, 7(4):29 - 31.

[3] 冯育麒. 车流径路域研究[J]. 铁道学报, 1996(4):20 - 24.

[4] 张　华. 特定经由标记语言的车流径路计算分析方法[J]. 铁道运输与经济, 2000(7):31 - 33.

[5] 铁道部. 铁路货物运价规则[M]. 北京:中国铁道出版社, 1996.

[6] 铁道部. 货运运价里程表[M]. 北京:中国铁道出版社, 1999.

(上接第 165 页)

参考文献:

[1] 陈文伟. 决策支持系统及其开发[M]. 北京:清华大学出版社, 2004.

[2] 戴劲松, 白英彩. 基于贝叶斯理论的垃圾邮件过滤技术[J]. 计算机应用与软件, 2006, 23(1):111 - 112.

[3] 胡　健, 马范援. 基于 Morphology 处理和主题词抽取的垃圾邮件过滤方法[J]. 上海交通大学学报, 2005, 39(12):1964 - 1965.

[4] 赵　伟, 戴新宇, 尹存燕, 等. 一种规则与统计相结合的汉语分词方法[J]. 计算机应用研究, 2004(3):23 - 25.

[5] He J, Tan A. On machine learning methods for Chinese documents classification[J]. Applied Intelligence, 2001, 18(3):311 - 322.