基于本体的电子邮件过滤策略

Tactic Of E-mail filtering based on Ontology

路 康 靳贺敏 （郑州轻工业学院电气信息工程学院 郑州）
杨 枫 （河南中医学院信息技术学院 河南郑州 450008）

摘要：针对各种邮件过滤方法的缺陷，本文提出了一种基于本体的邮件过滤方法。该方法利用对解码后邮件文档的元数据提取和本体标注，对其进行基于概率的 Naive Bayes 本体概念识别分类，并做出了语义解释和查询推理，从而实现了智能的邮件过滤。

关键词：过滤 语义网 元数据 本体

1 引言

电子邮件是 Internet 最重要的功能之一，它已经成为一种快捷、经济的通信手段，几乎所有连到 Internet 上的人都会使用电子邮件，但它带来方便的同时，也产生了大量的无用的垃圾信息。每天处理大量的垃圾邮件，对用户来说是件痛苦的事，这就需要能有一种自动对这类垃圾邮件进行过滤的处理方案和策略。

早在电子邮件开始使用的时候，人们就已经认识到垃圾邮件带来的危害，针对电子邮件的特点，各种组织和个人纷纷设计出了垃圾邮件的过滤方法。目前邮件过滤的方法有如下几种：

（1）基于邮件地址的过滤方法，包括地址过滤和安全认证。

地址过滤是指在地址拒绝列表中人工增加许多“黑名单”，地址包含在“黑名单”列表中的邮件将被删除或者拒绝接受。这种方法需要人工参与，但是垃圾邮件发送者往往假冒地址，如果垃圾邮件很多，这将给过滤带来难度。

安全认证方法，也就是用户 A 向用户 B 发送邮件时，必须到用户 B 的邮件服务器上先进行登记，得到授权，否则邮件服务器拒绝接收。这样虽然有效地防止未经认证的用户发来邮件，具有很高的安全性，但影响了邮件的易用性。

（2）基于邮件内容的过滤方法，包括规则判断、统计方法和语义分析

基于规则的方法很多时候是基于关键词匹配的邮件过滤，虽然能够处理邮件头和正文，但是实质还是生硬的二值判断，局限在二维空间上进行处理，因为缺少可信度方面的知识，同时要求用户自己定义规则，因此对用户的素质要求高，用户需要花费很多时间定义自己的规则，如果用户的兴趣发生变化，规则也要进行很大的改变，另外规则纯粹由人工定制，可能考虑并不周全。

统计的方法由于忽略了具体的语义环境，因此也只能区分合法邮件和垃圾邮件，很难进行深入分类。

语义分析方法主要对敏感的关键词出现的单句进行分析，依据主语、谓语、宾语的性质确定单句的表达含义。这在一定程度上可以对邮件进行判断，但是它是基于对单句的理解，以偏概全，往往把非垃圾邮件当成垃圾邮件来处理。

针对上述的邮件过滤方法，人们做了很多的尝试，但是都不能很好的解决垃圾邮件问题，其根本问题是计算机不能像人一样理解邮件内容。那么假如能够让计算机“读懂”邮件内容，让计算机理解邮件，进而对邮件进行推理判断，利用计算机的智能去过滤邮件，将能够彻底的避免邮件过滤的各种弊端，这是语义网技术研究的内容，也是本文的重点。

2 系统模型及原理

2.1 语义 Web 和本体

语义 Web 是现有的万维网的变革和延伸，它将使计算机能够“理解网上信息的含义”。在语义 Web 上，

信息都带有显式的含义，使其易于机器自动处理和Web信息集成。语义Web利用了XML可以自定义标签模式(tagging schemes)的能力和RDF可以灵活表示数据的能力。W3C组织提出了一些与语义Web有关的建议，包括XML、XML模式、RDF、RDF模式等。XML为结构化文档提供了基本的语法，但对文档的含义并未施加任何语义上的限制；XML模式是一种约束XML文档结构的语言；RDF是一个关于对象(或资源)和它们之间关系的数据模型，并为这个数据模型提供了简单的语义。这些数据模型使用XML语法表示。RDF模式是描述RDF资源中属性和类的词汇表，并带有这些属性和类的泛化层次的语义。这些建议构成了一个七层逻辑模型。本体层是该模型中重要的一层。

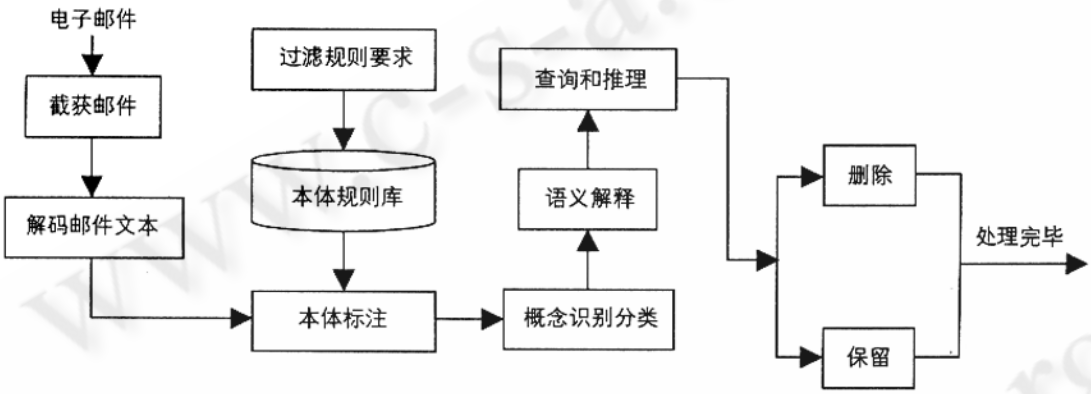

图 1　系统模型

本体(Ontology)是关于一些主题的清晰规范的说明。它是一个规范的、已经得到公认的描述，它包含词表(或名词表/术语表)。词表中的术语与某一领域相关，词表中的逻辑声明用来描述术语的含义和术语间关系。本体提供了用来表达和交流某些主题知识的词表和把握着词表中这些术语间的联系的关系集。一个本体包括一套关于某一领域概念的规范而清晰的描述称为类(class)或概念(concepts)，描述了有关概念的各种特性的属性(properties)和属性插件(slots)，还包括属性插件的限制条件(restrictions)和分面(facets)，以及一系列与某个类相关的实例(instances)，这些实例组成了一个知识库(knowledge base)。

2.2 系统模型及原理

本系统建立在对邮件本体的建立、理解、分类和推理之上的，由计算机根据用户的邮件过滤规则要求自动对邮件文档进行本体标注，然后利用基于概率的Naive Bayes分类器进行邮件的本体分类，进而用模型化的方法来进行语义解释，并做出查询和推理，从而分拣出垃圾邮件。

系统的模型如图1所示。主要由邮件拦截解码模块、本体标注模块，概念识别分类模块、语义解释模块和查询推理模块组成。

拦截解码模块运行在所需检测的电子邮件服务器上。为实现对电子邮件进行拦截，需要对电子邮件服务器端的软件进行修改，使其在转发邮件的时候自动调用过滤系统。同时识别邮件文本的编码类型，并进行相应的解码，以得到电子邮件信息。

本体标注模块是系统最主要的部分，它需要把解码后的电子邮件结果进行基于本体的元数据提取。整个提取过程是自动进行的，提取的数据主要是：邮件文档的本体分类，邮件文档的基本属性和与本体相关的属性，文档间的关系等。提取元数据之后需要对所得到的信息集进行本体标注，生成RDF描述。主要的流程如下：(1)对编码后的邮件文档进行预处理，主要是文本提取，去除对RDF元数据抽取无用的文本等；(2)进行基于本体的文档分类；(3)针对相应类别，进行属性提取，提取过程将使用相应的提取规则，这些规则主要是通过对样本集的分析和生成；(4)把生成的数据转化为RDF描述。

概念识别分类、语义解释和查询推理将在下文中详细描述。

3 系统分析与设计

3.1 邮件的本体概念识别分类

对邮件本体概念识别的过程也即成为一个分类的过程，与普通的基于内容主题的分类相比，基于本体的分类有其特殊的地方：(1)分类的主要依据不是邮件的主题内容，而是邮件的其他信息，这些信息将决定应该

使用本体中的什么概念去描述邮件；(2)分类还需综合考虑以后的本体属性和关系的识别。

因此，本系统决定使用基于概率的 Naive Bayes 分类器进行邮件的本体分类，文档的表示采用特征单词集(bag－of－words)的方式。这种分类器建立在这样的假设基础上的，即特征单词在文档中的出现是彼此独立的，尽管假设事实上不成立，但是它在实践中还是表现出很好的分类能力。

假设文档 d 的分类特征可由 $w_1, w_2, \cdots, w_n$ 个互相独立的单词表达，则根据 Bayes 规则，文档 d 属于分类 C_i 的概率为：

$$\begin{aligned}Pr(C_i|d) &= Pr(C_i|w_1, w_2, \cdots, w_n)\\ &= Pr(C_i)\prod_{i=1}^{n}\frac{Pr(w_i|C_i, w_1, w_2, \cdots w_n)}{Pr(w_i|w_1, w_2, \cdots, w_n)}\\ &= Pr(C_i)\prod_{i=1}^{n}\frac{Pr(w_i|c_i)}{Pr(w_i)} \propto Pr(C_i)\prod_{i=1}^{n}Pr(w_i|C_i)\end{aligned}$$

如果用 $|C_i|$ 表示类别 C_i 中所包含的样本文档数，用 $|D|$ 表示样本总文档数，则 $Pr(C_i)$ 可表示如下：

$$Pr(C_i) = \frac{|C_i|}{|D|}$$

用 $N(w_i, d_k)$ 表示特征单词 w_i 在文档 d_k 中出现的次数，用 $|V|$ 表示进行样本训练时所用的特征词汇总数，则 $Pr(w_i|C_i)$ 可通过如下公式计算得到：

$$Pr(w_i|C_i) = \frac{1 + \sum_{d_k \in C_i} N(w_i, d_k)}{|V| + \sum_{s=1}^{|V|}\sum_{d_k \in C_i} N(w_s, d_k)}$$

这样，就可以计算出带分类的邮件文档 d 属于各个类别的概率，而概率最大的那个分类就可以认为是该文档 d 所属的分类。

3.2 邮件本体的语义解释

我们用模型化的方法来描述邮件分类文档的语义。模型论假设这种语言是针对一个世界，并且描述了这个世界必须满足的最小条件集，从而给语言中的表达式指定相应的语义。一个特定的世界称为一个解释。这样模型论就可以看作是解释的理论。这个理论提供了一种抽象的数学方法来描述任何解释所具备的特征，而对真正的本质和内在结构作尽量少的假设。模型论试图中立于形而上学和本体论，尽量使用数学的集合论的知识作为基础。

RDF 是一种用规范的词汇表来表达命题(proposition)的断言(assertion)语言。确切地指出 RDF 断言的语义取决于许多因素，包括描述约定、自然语言的简述、与其他内容的链接等。RDF 中的每个三元组表达了一个命题，这给 RDF 定义下了一条相当严格的单调性质，从而使它不能表达封闭世界的假设、局部缺省的条件，以及其他非单调所常用的构造手段。

RDF 使用 URI(统一资源标识符)引用和常量指向表达式。以下用数学语言来定义一个解释 I。它相对于一个 URI 引用集而言，称为词汇表。

VL：所有常量值的集合，这个集合包括所有的常量字符串，<常量字符串，界标>对的集合，以及所有规范的 XML 文本；

R：非空资源集，称为 I 的论域或定义域，R 是 VL 的超集；

P：R 的子集，包含所有的性质；

IEXT：从 P 到 R 幂集的映射，使得 IEXT(p)＝{<r1, r2>|r1 和 r2 属于 R}，p 属于 P；

IS：从 V 到 R 的映射，使得 IS(v)＝r，r 属于 R，v 属于 V；

IL：从 VT 到 R 的映射，使得 IL(v')＝r，r 属于 R，v' 属于 VT。

说明：IEXT(x)，x 属于 P，是一个对的集合。集合中的每一个元素都标定了两个参量，使得 x 为真。这两个参量也可以称为二元关系外延，即 x 的外延。

解释 I 中基本图的指称语义由以下规则递归的给出。首先定义 RDF 表达式 E 为 URI 引用或常量。

如果 E 是简单常量，则 I(E)＝E；

如果 E 是有类型常量，则 I(E)＝IL(E)；

如果 E 是一个 URI 引用，则 I(E)＝IS(E)；

如果 E 是一个三元组 <S,P,O>，当 <I(S), I(O)> 属于 IEXT(I(P)) 时，I(E) 为真，否则 I(E) 为假；

如果 E 是一个基本 RDF 图，则当 I(e') 为假时，I(E) 为假，否则 I(E) 为真，e' 是 E 中的一个三元组。

注意简单常量的指称语义总是在 VL 中。

如果一个 RDF 图 G 的词汇表 V' 包含某个 URI 引用 u，而 u 未出现在解释 I 的词汇表 V 中，即未能给出 G 中的名称 u 的语义，则以上规则作用在 u 上使得 I(u) 为假，从而使得 I(G) 为假。反过来说，一个图的断言认定图中的名称实际上是指向世界中的某些事物。这个条件隐含着一个空图为真。

举一个例子。设一个语义解释 I 定义如下：

词汇表 V = {ex：a，ex：b ex：c}

R 为 VL 与{1,2}的并，用整数表示论域(特定领域)中的非常量事物；

IEXT：1 - - >{ <2,3 >，<3,2 >}

IS：ex：a - - >1，ex：b - - >2，ex：c - - >3；

IL：VT - - >2.

以上解释 I 满足下列三元组：

<ex：a > <ex：b > <ex：c >

<ex：c > <ex：a > <ex：b >

"xx" <ex：a > <ex：c >

<ex：c > <ex：a > "yy"

但是不满足下列三元组：

<ex：a > <ex：b > <ex：c >

<ex：a > <ex：c > <ex：b >

<ex：b > <ex：c > <ex：a >

3.3 查询和推理

对经过本体标注并分类和语义解释后的电子邮件文档，需要进行逻辑上的查询推理和判断，从而决定哪些是垃圾邮件，进而将其删除。

为了能够对 RDF 所表述的语义解释进一步的推理，需要有一个通用的标准语言来表示本体。下面将利用 W3C 提出的 OWL(Web Ontology Language)语言来做进一步的分析。

OWL 相对 XML、RDF 和 RDF Schema 拥有更多的机制来表达语义，它能够被用于清晰地表达词汇表中的词条(term)的含义以及这些词条之间的关系。对 OWL 而言，语义逻辑的处理才是推理机制的实现。本系统利用 jena2.1 提供的 OWL 接口。

假设我们给出这样两个文件：owlDemoSchema.owl 和 owlDemoData.rdf，前者给出了一个关于邮件的本体，它将定义这样一类邮件：必须至少包含一个发件人(FromMan)以及一个值为张三(zhangsan)的组件。owlDemoData.rdf 文件定义了一些假定的邮件属性和相应的特征描述。

可以利用以上两个关联文件创建一个推理器(Reasoner)实例，代码如下所示：

```
Model schema = ModelLoader.loadModel("file：owlDemoSchema.owl")；
Model data = ModelLoader.loadModel("file：owlDemoData.rdf")；
Reasoner reasoner = ReasonerRegistry.getOWLReasoner()；
Reasoner = reasoner.bindSchema(schema)；
InfModel infmodel = ModelFactory.createInfModel(reasoner,data)；
```

对以上模型的一个典型操作就是查找特定实例，如查找一个广告(advertisement)邮件。实现代码如下：

```
Resource advertisement = infmodel.getResource("urn：x - hp：eg/ advertisement")；
System.out.println("advertisement ＊：")；
PrintStatements (infmodel, advertisement, null,null)；
```

3.4 系统实现思想

当邮件服务器收到新邮件后，首先对新邮件进行解码，获取邮件主要内容，在由过滤规则产生的过滤本体的协助下对其进行标注和结构化编码，变无结构的数据为有结构的数据，然后由基于概率的 Naive Bayes 分类器对新邮件进行分类，从而判断出新邮件属于哪一类别，并获取新邮件的主要概念，接着利用过滤本体和邮件主要概念对新邮件进行语义解释，使得计算机有能力对其进行推理，最后对新邮件进行逻辑上的查询推理和判断，从而决定哪些是垃圾邮件，并对其进行处理。

对新邮件进行解码的目的是为了抽取新邮件的元数据，其解码过程目前有多种成熟的技术。通过解码不仅能对邮件进行初步的筛选剔除，而且方便过滤本体对其进行标注和编码。对新邮件的分类、语义解释以及查询推理上文已经作了简单的阐述，这里需要说明的是如何对新邮件进行语义标注和结构化编码。

语义标注是根据有关概念集为文档及其各个部分标注概念类、概念属性和其他元数据的过程，是语义推理的基础。利用过滤本体建立词汇集合与概念类别之间的映射关系，然后通过自动词汇分析找出邮件内容的概念类别，甚至与其他类别的语义关系，利用这些概念类别进行标注。语义标注最终要完成利用 RDF 对邮件内容进行描述编码的任务。假如邮件内容里有这样的一句话"IEEE 邀请参加 2006 年网络年会"，那么利用过滤本体进行语义标注和 RDF 编码的结果为：

(下转第 46 页)

(上接第 35 页)

<rdf:Description rdf:about = "http://www.untartyu.com/2006 年网络年会" >

<dc:邀请参加 > IEEE </dc:邀请参加 >

</rdf:Description >

这样就将无结构的数据内容变成用 RDF 描述的结构化语句,方便进行处理。

4 结论

电子邮件过滤技术是 Internet 上各种信息处理过滤的一个组成部分。本文分析了当前各种邮件过滤技术存在的缺陷,并给出了一种基于本体的电子邮件过滤策略,将从根本上改善了垃圾邮件的过滤效果。由于目前语义网技术发展还处在开始阶段,对本体推理技术的研究还不太成熟,所以本技术距离实用还有很多的工作要做,并且还有很多需要改进的地方,但是随着研究的进一步深入,邮件过滤技术也会进入一个更高的层次。

参考文献

1 M. Erdmann , R. Studer. How to Structure and Access XML Documents with Ontologies. Data and Knowledge Eng. ,2001 ;36 (3).

2 Henry Kim. Ontology Applications and Design : Prediction How Ontologies for the Semantic Web Will Evolve ,Communications of ACM ,2002 ;45(2).

3 John M. Pierre. On the Automated Classification of Web Sites. L Inkping Elec2 tronic Articles in Computer and Information Science ,2001 ;(6).

4 陈华辉,一种基于潜在语义索引的"垃圾"邮件过滤方法[J],计算机应用研究,2000,10.

5 Jason D, Rennie M. An Application of Machine Learning to E – mailFiltering. Sixth ACM SIGKDD International Conference on KnowledgeDiscovery and Data Mining,Boston,MA,USA,2000 – 08 – 20.