广西科学院学报
Journal of Guangxi Academy of Sciences

2008,24(4):310~312,316
Vol. 24,No. 4 November 2008

基于 Web 挖掘自动分类器的设计与实现*
Design and Implementation of the Auto Classifier Based on Web Mining

朱春江,陆宇昱,李陶深,杜衡斌,唐 晟
ZHU Chun-jiang,LU Yu-min,LI Tao-shen,DU Heng-bin,TANG Sheng

(广西大学计算机与电子信息学院,广西南宁 530004)
(School of Computer, Electronics and Information, Guangxi University, Nanning, Guangxi, 530004,China)

摘要:分析分布式实时网络行为监控系统中 Web 网页安全性挖掘问题,设计实现一个基于 Web 挖掘的自动分类器,并构造一个实验环境来检测分类器的性能。该自动分类器利用特征提取算法实现对每个样本的特征向量提取和待分类文本的特征向量提取,利用基于 k 个"最近邻"(KNN)分类算法实现对网页的分类,能够提取出带有不安全信息的网页,分类效果良好。

关键词:网络行为监控 Web 网页挖掘 分类器 KNN 分类算法 特征提取
中图法分类号:TP393.092 **文献标识码**:A **文章编号**:1002-7378(2008)04-0310-03

Abstract: This paper analyzes Web security mining problem in distributed real-time network behavior monitoring system. An auto classifier based on Web minning was designed and implemented. An experiment environment to test the performance of the classifier was constructed. This classfier extracts the feature vector of each samples and documents to be classified by using the feature extraction algorithm. Web page was classfied by using the K-Nearest-Neighbor(KNN) classification algorithm. The experimental results show that this auto classifier based on Web minning can fetch insecurity Web pages, and its classification is effective.

Key words: network behavior monitoring, Web page minning, classifier, KNN classification algorithm, feature extraction

对于面向 Internet 的分布式网络监控系统来说,行为监控的最终目的是实施管理。如果被监控用户的行为有害或违反规则,管理者应该能够发现它的有害行为,进而采取一定的限制策略来对其限制,例如发出警告、限制网速或关闭端口。然而,Internet 网作为一个巨大、广泛分布的全球信息中心,其内容信息涉及新闻、广告、消费者信息、财务管理、教育、政府、电子商务等,此外还包含了丰富和动态的超连接信息及访问和使用信息。在这些信息中,带有不安全因素的网页或网络行为只是这海量数据中的小部分。考虑到监控的实时性,必须利用数据挖掘方法来实现数据的监控[1]。

网页内容挖掘的对象包括文本、图像、音频、视频、多媒体和其他各种类型的数据。网页内容挖掘一般从资源查找和数据库两个方面来进行研究[2]。从资源查找(IR)的观点来看,网页内容挖掘的任务是从用户的角度出发,怎样提高信息质量和帮助用户过滤信息。而从数据库的角度讲网页内容挖掘的任务主要是试图对网页上的数据进行集成、建模,以支持对网页数据的复杂查询。怎样从海量监控的网页信息中提取出带有不安全因素的网页数据,进而采取相应的响应策略,是开发一个具有高效率的分布式实时网络监控系统必须考虑的因素。本文结合网络行为监控系统的目的,设计一个基于 Web 挖掘的自动分类器,实现对网页内容的监控。

1 自动分类器流程设计

1.1 流程分析

首先,网络行为监控系统要实施人性化的管理,

收稿日期:2008-10-12
作者简介:李陶深(1957-),男,教授,主要从事分布式数据库系统、计算机网络技术等研究。
*广西科技攻关项目(桂科攻关 033008-9)资助。

必须检测出每个用户的行为的危害性。行为的危害性可以简单的理解为用户所浏览网页的安全性。因此,拓展中重点考虑用户浏览网页的安全性,即网页页面安全性挖掘。

其次,网络行为监控系统工作时,监控网络中被监控计算机的数量可能很多,而每一个被监控计算机浏览的网页更可能有很多个网页。所以,对于捕获到的大量网页,需要经过还原之后,再对其进行基于 Web 挖掘的处理。为了满足实时性要求,基于 Web 挖掘的处理应该设计比较简单,耗时要少。

再次,分析用户所浏览网页的安全性,可以考虑简单地对网页进行分类。例如,可以简单地将网页分成两类,一类是带有不安全信息的网页,另一类则是安全的网页。当然,分类之前必须收集两类各自的样本。分类是基于文本来分,分类之前还必须提取文本的特征,然后才能根据其特征进行分类。

1.2 流程设计

根据分布式实时网络行为监控系统中 Web 网页安全性挖掘问题的特点,基于 Web 挖掘的自动分类器设计有网页预处理、中文分词、特征提取和分类 4 个功能模块,其处理流程设计如图 1 所示。

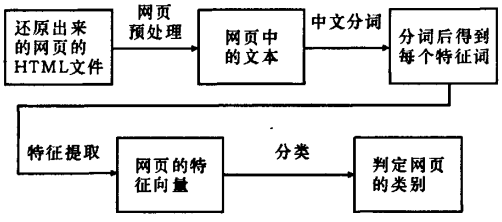

图 1 基于 Web 挖掘的自动分类器处理流程

在基于 Web 挖掘的自动分类器处理流程中,网页预处理的任务是把 HTML 文件中的文本信息提取出来。此处的"HTML 文件中的文本信息"是指网页在显示时用户所能看到的所有文字信息。网页预处理提取出来的文本信息是否完整将直接影响到随后分类的结果的正确性,因此它是分类器设计的关键技术。

中文分词阶段的任务是对预处理阶段提取出来的 Web 页面主要信息进行分词,得到一个个的词及词性。我们采用基于多层隐马尔可夫模型的汉语词法分析系统(ICTCLAS)来实现中文分词工作,该系统的分词正确率高达 97.58%[3]。

经过中文分词之后,HTML 文档已经转换成为一个个独立的词及其词性。这些词可以理解成为文本的特征。综合考虑网页文本特征提取的特征,我们选取采用向量空间模型对文本进行结构化处理,以进行网页文本特征提取。

作为 Web 挖掘工作的最后一步,分类的任务是要构造一个模型或分类器来预测类标号。我们使用分类效果较好且实现方法比较简单的 KNN(K 最近邻)方法[4]实现分类。分类过程分为训练过程和分类过程。使用训练数据进行训练之后,发现数据之中蕴含的模式,利用发现的模式对待分类数据进行分类。

2. 自动分类器实现的关键技术

2.1 特征提取算法

使用向量空间模型对经过中文分词之后的文本进行结构化处理,然后将某个文本的特征向量提取出来。自动分类器中的特征提取算法的流程如图 2 所示。

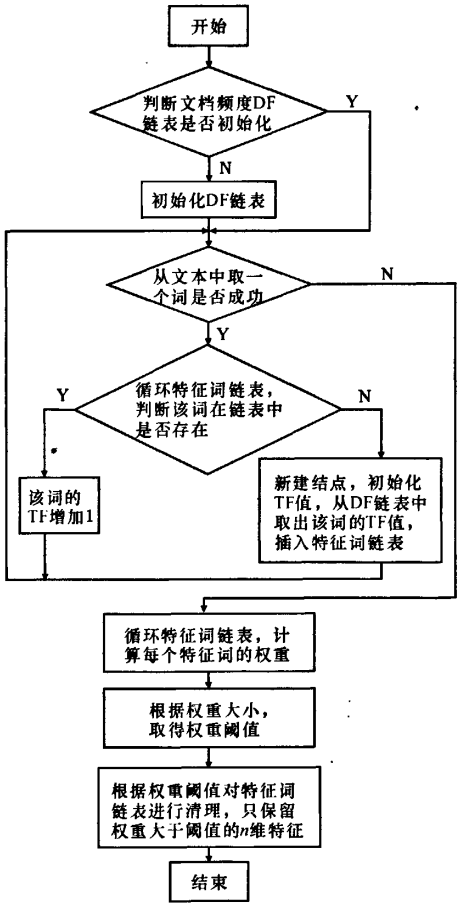

图 2 特征提取算法流程

2.1.1 从文本中取一个词

判断"从文本中取一个词是否成功"是一个比较复杂的过程,它的任务是从文本中取出一个词放到

记录当前词的全局数据结构CurWord中。成功表示取到一个词，失败表示此文本中词已经取完。在特征提取算法中，缓冲区不管取多大，都不能保证一个特征词不会被它的边界所打断。为了解决这个问题，缓冲区采用"一分为二"的方法。即设立两个互补使用的缓冲区，当一个缓冲区使用完毕，则将新的数据装入另一个缓冲区，装入之前必须把上次没有处理完的缓冲区尾部的字符全部拷贝至新的缓冲区的首部，装载完上次没处理完的字符后再从文件中提取新的字符填满缓冲区。这种方法可以很好的解决缓冲区边界的问题，但是要进行限制，即没有处理完的字符的长度不能大于一个缓冲区的长度。如果不加以限制，设定再大的缓冲区也无济于事。

2.1.2 特征提取的评估方法

在向量空间模型中，特征提取的评估方法采用文档频数DF权值方法，权值用TFIDF公式 $w_{ik}=tf_{ik}\times idf_k$ 计算，其中，tf_{ik} 表示特征词 t_k 在文档 d_i 中出现的频率，idf_k 表示该特征的反比文本频数。它表示在一个文本中出现次数越多的词，在另一个同类文本中出现的次数也会很多，反之亦然。该方法是根据特征词的重要性与特征词的文档内频数成正比，与训练文档中出现该词条的文档频数成反比的原理构造的。

在基于Web挖掘的自动分类器中，我们对TFIDF公式进行改造，采用公式 $w_{ik}=tf_{ik}\times\log(N/n_k+0.01)$ 来计算每个词的权重值，其中，N 表示全部训练库中的文本数量，n_k 则表示训练文本中出现 t_k 的文本数。例如，训练库文本总数为5000篇，在一篇文档中特征词"广西大学"出现了2次，因此它的 t_f 值为2；在所有训练库中有50篇中出现了"广西大学"这个词，因此，"广西大学"这个词在该篇文档中的权重值 $w_{ik}=2\times\log(5000/50+0.01)=4$。

划定权重阈值根据要提取特征向量的维数来决定。当设定特征向量的维数是50，则计算阈值，按照权重值从大到小的排列，使得阈值等于权重值排名第50名的特征词的权重。计算出权重的阈值以后，最后一步要对特征词链表进行清理，保留权重大于等于阈值的特征词，此处大于等于阈值的特征词可能不止设定的维数的大小，因为可能有很多词的权重值是相等的，此时只需提取出维数大小的特征向量后，删除掉剩下的所有特征词。

经过特征提取算法的处理后，一篇文本的特征向量就被成功提取。

2.2 分类算法

KNN算法的基本思想是：基于类比学习，即通过给定的检验元组与和它相似的训练元组进行比较来学习。训练元组用 n 个属性描述。每个元组代表 n 维空间中的一个点。这样，所有的训练元组都存放在 n 维模式空间中。当给定一个未知元组时，k 最近邻分类法搜索该模式空间，找出最接近未知元组的 k 个训练元组。这 k 个训练元组是未知元组的 k 个"最近邻"。未知元组指派到它的 k 个最近邻中的多数类[1]。

将KNN算法应用在基于Web挖掘的自动分类器中的思路是：在给定待分类文本后，考虑在样本文本集中与该待分类文本距离最近(最相似)的 K 篇样本，根据这 K 篇样本所属的类别判断待分类文本所属的类别。分类算法的流程如图3所示。

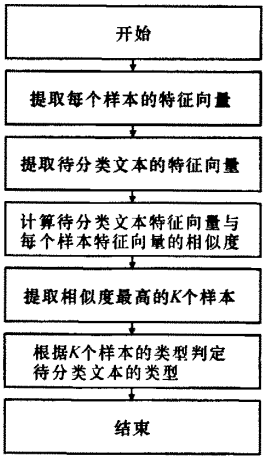

图3 分类算法流程

在基于Web挖掘的自动分类器处理流程中，利用特征提取算法实现对每个样本的特征向量提取和待分类文本的特征向量提取，相似度则使用余弦距离公式[4]进行计算；在"根据 K 个样本的类型判定待分类文本的类型"时，可以利用公式

$$p(x,C_j)=\sum_{i\in KNN}Sim(x,S_i)y(S_i,C_j)$$

计算每一类的权重，其中，x 为待分类文本的特征向量，$p(x,C_j)$ 为此文本属于类 C_j 的权重，S_i 表示第 i 个样本，$Sim(x,S_i)$ 表示利用余弦距离公式算出的待分类文本的特征向量与第 i 个样本的特征向量的相似度，$y(S_i,C_j)$ 表示判别函数，即如果第 i 个样本属于类 C_j，则其值为1；否则为0。算出的权重值是待分类文本属于每一类的权重；选取权重最大的类，判断待分类文本属于该类。

(下转第316页)

示在Linux平台下通过对目标程序连接后生成了可执行程序及信息输出，同样采用中英文对照显示，使编程者易于理解。Windows平台下也具有类似的界面。无论在Windows或Linux环境下，系统都能新建、打开、编辑、保存各种不同的文件，包括ASM、TXT、TEXT等类型或无类型的文件。

4 结束语

随着Linux应用的越加广泛，采用Windows、Linux双系统兼容设计的软件将会越来越多，本文设计实现的跨平台汇编语言集成系统，能够兼容Windows和Linux系统平台，该系统不仅有利于工程技术开发，也有利于初学者掌握汇编语言程序的开发过程。目前该系统已经在汇编语言教学中应用，它以其方便实用的特点得到了普遍欢迎。

参考文献：

[1] 刘跃华，梁英. 智能化的汇编语言集成开发环境[J]. 计算机技术与自动化，2007，26(3)：114-117.

[2] 沈洁萍，章红，陈勇. 汇编语言集成编译环境的开发[J]. 微计算机信息，2004，20(6)：120-121.

[3] 蔡启先，王智文，黄晓璐. 汇编语言程序设计实验指导[M]. 北京：清华大学出版社，2008.

[4] 陈火旺，刘春，谭庆平，等. 程序设计语言编译原理[M]. 第3版. 北京：国防工业出版社，2000.

[5] 耿祥义，张跃平. Java2实用教程[M]. 北京：清华大学出版社，2006.

[6] 马季兰，彭新光. Linux操作系统[M]. 北京：电子工业出版社，2002.

（责任编辑：韦廷宗）

（上接第312页）

3 实验分析

构造一个简单的样本集，样本由60个网页组成。使用30个与NBA有关信息相关的网页构成前30个样本，使用另外30个与NBA不相关的网页构成后30个样本。类别分成两类，一类是NBA类，一类是其他类。利用KNN分类程序对待分类文本进行分类，观察分类结果。

在60个待分类文本中，NBA类的准确率是90%。其他类的准确率是76.67%。分类时间大约3s。在实验环境下，样本中的NBA类可以看作是“带有不安全信息”的网页类，其他类则看作是安全的网页类。能够正确分析出某一待分类网页是否为NBA类，则说明分类器的工作是有效的。

4 结束语

本文对网页的安全性进行挖掘，设计实现一个基于Web挖掘的自动分类器，并构造一个实验环境来检测分类器的性能。实验结果表明，所设计的自动分类器可以得到比较好的分类结果。但是由于KNN属于惰性分类方法，所以耗费的资源较大，速度相对较慢。我们下一步的工作将继续研究Web挖掘中关于网页安全性的判断和检测方法，进一步提高自动分类器的分类准确率，在准确率和速度之间寻求平衡点。

参考文献：

[1] Han Jiawei, Micheline Kamber. 数据挖掘：概念与技术[M]. 第2版. 北京：机械工业出版社，2007.

[2] 冯迪，李晋宏，曹原. 基于网页的数据挖掘研究：2007通信理论与技术新发展——第十二届全国青年通信学术会议论文集：上册[C]. 北京：电子工业出版社，2007.

[3] 张华平，刘群. 基于N-最短路径方法的中文词语粗分模型[J]. 中文信息学报，2001 (5)：1-7.

[4] 梁循. 数据挖掘算法与应用[M]. 北京：北京大学出版社，2006.

（责任编辑：韦廷宗）