文章编号:1001 - 9081(2006)10 - 2427 - 03

Web 浏览器历史数据自动分类取证系统

石淼磊[1],苏璞睿[2],冯登国[2]
(1. 中国科学技术大学 电子工程与信息科学系,安徽 合肥 230026;
2. 中国科学院软件研究所 信息安全国家重点实验室,北京 100049)
(sml@ustc.edu)

摘 要:为提高取证的自动化程度,提出了一种基于页面自动分类技术的浏览器历史数据取证算法,并设计实现了一个原型系统。该系统在获取浏览器历史数据的基础上,自动对其进行特征提取、页面分类。实验结果表明该系统有效提高了取证人员的效率和准确度。

关键词:Web 浏览器;计算机取证;页面分类;数据挖掘
中图分类号: TP309.2 **文献标识码**:A

Automated categorization forensic system for history data of Web browsers

SHI Miao-lei[1], SU Pu-rui[2], FENG Deng-guo[2]
(1. *Department of Electronic Engineering and Information Science, University of Science and Technology of China, Hefei Anhui* 230026, *China;*
2. *State Key Laboratory of Information Security, Chinese Academy of Sciences, Beijing* 100049, *China*)

Abstract: To enhance the automation of forensics, an automated categorization forensic method for history data of Web browsers based on Web classification technology was proposed in this paper, and a prototype system was implemented. The system automatically extracted the features of history data of web browsers and categorized the caught Web pages. The experimental results show that the system greatly increases the forensic efficiency and accuracy.

Key words: Web browser; forensic; categorization; data mining

0 引言

随着网络的高速发展,利用计算机和网络从事非法活动或交易的犯罪活动也越来越猖獗,对计算机犯罪活动的取证已成为安全领域的重要研究内容。Web 浏览器作为一种广泛使用的网络信息发布和获取工具,其历史数据中包含了重要信息,比如历史访问 URL、Cookie、缓存页面等。目前在对各类浏览器历史数据分析的基础上,一些公司和组织都发布了自己的工具,其中 Free Stone 发布的 Pasco,支持 Internet Explorer 取证;MANDIANT 公司发布的 Web Historian 和 DataTex Engineering 发布的 Forensic Tool Kit,支持 Internet Explorer、Mozilla firefox、Opera 等大部分浏览器的取证。但是这些取证工具都局限于 Web 浏览器历史数据的存储格式的分析,只提供了历史 URL 列表、Cookie、缓存页面等基本信息,仍需取证人员在复杂的信息中人工地分析出电子证据,这个操作通常很费时费力,而且出错率比较高。

本文在前人工作的基础上,设计并实现了浏览器历史数据自动分类取证系统(Automated Categorization Forensic System for History Data of Web Browsers,ACFS),它在对浏览器历史数据格式进行分析的基础上,利用数据挖掘技术根据页面内容进行自动分类处理,进一步提高了取证的自动化程度,减轻了取证人员的负担。

1 系统总体架构

ACFS 系统是一个基于自动分类算法的浏览器历史数据取证系统。它具有标准接口,易于扩展为针对其他数据取证的自动分类取证系统,其系统结构如图 1 所示。

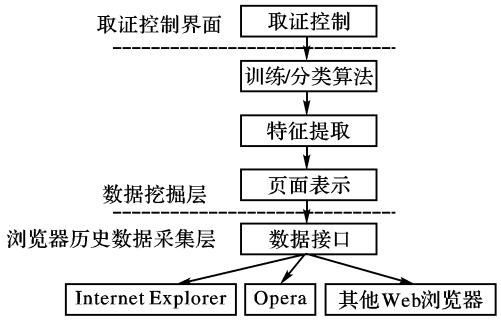

图 1 ACFS 系统结构图

整个系统分为三个层次,分别为:

1)取证控制界面:主要用于系统管理,包括浏览器类型配置、历史数据路径配置,自动分类结果处理。

2)数据挖掘层:主要包括三个部分,页面表示、特征提取、训练和分类算法。

3)浏览器历史数据采集层:主要实现浏览器的历史数据的采集,包括访问页面的 URL、Cookie、缓存页面内容等的提取,并且提供统一的数据接口给数据挖掘层。

其中取证控制界面主要实现与用户的交互,不对其进行详细讨论。下面分别介绍浏览器历史数据采集层和数据挖掘层的具体实现。

2 浏览器历史数据采集层

浏览器历史数据采集层主要实现浏览器历史数据的采

收稿日期:2006 - 04 - 28;**修订日期**:2006 - 06 - 09

作者简介:石淼磊(1980 -),男,浙江绍兴人,硕士研究生,主要研究方向:信息安全; 苏璞睿(1976 -),男,湖北人,博士,主要研究方向:入侵检测、评估; 冯登国(1965 -),男,陕西人,研究员,博士,主要研究方向:信息和网络安全.

集,如访问页面的 URL、Cookie、缓存页面等。在 ACFS 系统设计的时候,考虑到系统的扩展性,以支持新的浏览器,本层主要分成数据接口和浏览器实现两个部分,如图 2 所示。

数据接口部分主要向上隐藏具体的浏览器数据采集实现,使数据挖掘层不需要考虑底层实现,而只需要调用公用的接口来提取数据。本接口主要提供给数据挖掘层三个方法(setBrowser、getURLs、getPage),分别用来配置浏览器属性,获取页面列表和获取页面内容。本接口将根据浏览器的属性来调用具体的浏览器实现部分的对应方法。

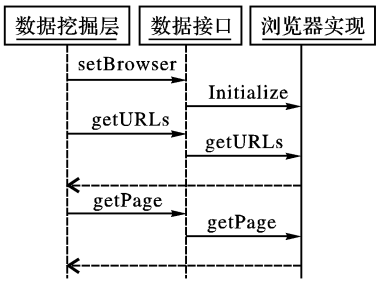

图 2 浏览器数据挖掘层框架图

浏览器实现部分主要实现各种浏览器的数据采集。因为不仅仅在不同的浏览器之间,即使在同一浏览器的不同版本之间,它们的历史数据格式、缓存页面存放方法等都有差异,所以对不同的浏览器的数据采集需要使用对应的方法。

目前 ACFS 系统主要实现了 Internet Explorer(版本 5 - 6)[1]、Mozilla Firefox(版本 1 - 1.5)和 Opera(版本 7 - 9)的历史数据采集。

3 数据挖掘

数据挖掘层主要对将从浏览器历史数据采集层获取的数据进行数据挖掘处理,实现页面自动分类。页面自动分类主要包括训练和分类两个阶段,每个阶段主要由页面表示、页面特征提取和实施分类算法组成,如图 3 所示。

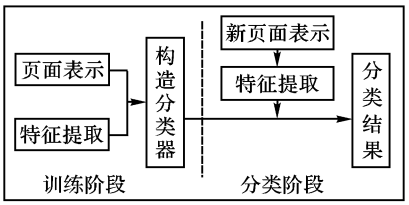

图 3 页面自动分类结构图

3.1 页面表示

由于计算机并不具有人类的智能,人在阅读文章后,根据自身的理解能力可以产生对文章内容的模糊认识,而计算机并不能轻易地“读懂”文章,所以必须将页面转换为计算机可以识别的格式。目前 Web 页面的表示主要采用向量空间模型(VSM)[2],即用一组特征项(t_{i1},t_{i2},t_{i3},…,t_{im})和相应的权重(w_{i1},w_{i2},w_{i3},…,w_{im})来表示相应的页面 d_i,其中 t_{ij} 是 d_i 的第 j 个特征,w_{ij} 是 t_{ij} 的权重。

3.1.1 特征项

通常选择页面中的词条作为特征项 t_{ij},对英文而言,由于英文的词与词之间有固定的分隔符,所以只需要进行 stemming 处理;但是中文的情况则不同,因为中文的词与词之间没有固定的切分标志,需要进行中文分词。在中文信息处理领域,中文分词主要分为基于字典、基于理解和基于统计三大类[3]。在 ACFS 系统中,采用了基于字典的逆向最大匹配算法,并且结合 Web 页面的结构特点,在分词的同时对词条在页面中的重要性进行标记,重要性设定如表 1 所示。例如,如果某词条在出现在 Title 标签里,则该词条的重要性标记为 6;如果出现在在普通正文部分,则其重要性标记为 1。

在对页面及词条分别进行分词和重要性标记后,页面 d_i 可以表示为((t_{i1},χ_{i1}),(t_{i2},χ_{i2}),(t_{i3},χ_{i3}),…,(t_{in},χ_{in})),其中 t_{ij} 是页面 d_i 的第 j 个词条,χ_{ij} 是 t_{ij} 在页面 d_i 的重要性。

表 1 词条在页面中的重要性

词条在页面中位置	重要性
页面标题(Title)	6
页面 Meta(Description, Keyword)	5
H1、H2、H3…	4
链接	3
Italic、Strong	2
普通正文	1

3.1.2 权重

在 VSM 中,权重用词频表示。词频分为绝对词频和相对词频,绝对词频,即词在页面中出现的频率,相对词频为归一化的词频,其计算方法主要运用 TF-IDF[4] 公式,下面是一个常用的 TF-IDF 公式:

$$w_{ij}=\frac{ft_{ij}*\log\left(\frac{N}{nt_{ij}}+0.1\right)}{\sqrt{\sum_{i=1}^{n}\left(ft_{ij}*\log\left(\frac{N}{nt_{ij}}\right)+0.1\right)^2}}$$

其中 N 为训练页面的数目,ft_{ij} 是词条 t_{ij} 在文档 d_i 中的词频,nt_{ij} 为含有词条 t_{ij} 的训练页面数。我们对其进行了改进,将词条在页面中的重要性引入 TF-IDF 公式,修正后的公式如下:

$$w_{ij}=\frac{\chi_{ij}*ft_{ij}*\log\left(\frac{N}{nt_{ij}}+0.1\right)}{\sqrt{\sum_{i=1}^{n}\left(\chi_{ij}*ft_{ij}*\log\left(\frac{N}{nt_{ij}}\right)+0.1\right)^2}}$$

从上面的 TF-IDF 公式可看出:

1)页面集合中包含某一词条的文档越多,说明该词条区分页面类别属性的能力越低,其权重 w_{ij} 越小;另一方面,某一页面中某一词条出现的频率越高,说明它区分页面内容属性的能力越强,其权重 w_{ij} 越大。

2)某词条在大部分页面的重要性 χ_{ij} 都很高的话,则它的重要性 χ_{ij} 通过平均计算将被抑制,权重 w_{ij} 会变小;如果某词条在少数的页面中的重要性 χ_{ij} 比较高的话,则它的重要性将被体现,权重 w_{ij} 会变大。

修正后的权重计算公式有效的结合了 Web 页面的结构特性,进一步体现了 Web 页面的特点。

3.2 特征提取

由于构成页面的词条的数量较多,导致表示页面的 VSM 的维数将相当大,即使对于中等大小的页面也可以达到几万维,因此需要在尽量不影响分类的前提下减少 VSM 的维数,即降维运算。这样做的目的主要包括两方面:第一,提高程序的运行效率;第二,提高分类精度,当前存在多种特征提取算法,比如:

1)根据文档频次(DF):去除那些文档频次特别低(太少,没有代表性)和特别高的特征项(太多,没有区分度)。

2)根据信息增益(Information Gain,IG):即根据特征项为整个分类所能提供的信息量。

3)根据互信息(Mutual Information,MI):即根据特征项和类别的相关性衡量特征项。

4)根据开方拟合检验(CHI,$\chi 2$ - test):即根据特征项和类别之间的独立性。

5)根据词熵判断、根据 KL(Kullback-Leibler divergence)距离等算法。

在 ACFS 系统中采用 kNN(k-Nearest Neighbor)分类算法

对 IE 浏览器的 1 032 个历史页面分别使用 DF、IG、MI、CHI 特征提取算法进行对照实验,得出如下结果:

表 3　特征分类结果精确度(kNN 分类算法)

精确度\算法	DF	IG	MI	CHI
精确度	0.897	0.912	0.825	0.904

从表 3 可以看出 DF、IG 和 CHI 的性能大体相等,MI 相对来说就差了,这个结果和文献[5]的结果大体一致。根据此结果,ACFS 系统将 IG 作为默认的特征提取算法。

3.3　训练方法和分类算法

训练方法和分类算法是 ACFS 系统的核心部分,目前存在多种基于向量空间模型的训练算法和分类算法[6]。在 ACFS 系统中,我们实现了简单向量距离、kNN[7] 和类中心分类这三种算法,其中 kNN 算法是效果最好的。

kNN 算法的基本思路是:在给定新页面后,考虑在训练页面集合中与该页面距离最近的 K 个页面,根据这 K 篇页面所属的类别判定新页面所属的类别,具体的算法步骤如下:

1) 使用 VSM 来表示训练页面集合;

2) 对新页面进行页面表示;

3) 在训练页面集中选出与新页面最相似的 K 个页面,使用的相似度计算公式如下:

$$Sim(d_i,d_j) = \frac{\sum_{k=1}^{M} w_{ik} \times w_{jk}}{\sqrt{(\sum_{k=1}^{M} w_{ik}^2)(\sum_{k=1}^{M} w_{jk}^2)}}$$

在 ACFS 系统中设 K 为 100。

4) 在选择的 K 个邻居中,依次计算属于每类的权重,公式如下:

$p(\vec{x},C_j) = \sum_{\vec{d}_i \in kNN} Sim(\vec{x},\vec{d}_i)\phi(\vec{d}_i,C_j)$ 其中,$\vec{x}$ 为新页面的特征向量,$Sim(\vec{x},\vec{d}_i)$ 为第 3 步的相似度计算公式,而 $\phi(\vec{d}_i,C_j)$ 为类别属性函数,即,如果 $\vec{d}_i$ 属于类 C_j,那么函数值为 1,否则为 0。

5) 比较类的权重,将页面分到权重最大的那个类别中。

除此以外,贝叶斯、支持向量机(SVM)和神经网络算法在页面分类系统中应用得也较为广泛。基于 ACFS 系统的设计框架,可以很方便地在今后的工作中引入对其他算法的支持。

4　实验数据分析

在实验中,我们将页面分成了色情、非法交易、反动言论和正常页面这 4 个分类,为每个类别分别选取了 300 个页面作为训练页面集合。然后在一台安装有 Internet Explorer 6.0 的 PC 机上进行取证分析,在获得的 1 032 个历史页面中进行自动分类取证,得到如下的实验结果:

表 4　实验结果

分类法\查准率	简单向量距离	kNN	类中心分类
正常页面	0.862	0.908	0.895
色情	0.908	0.926	0.918
非法交易	0.843	0.890	0.872
反动言论	0.915	0.932	0.929

从表 4 可以看出:kNN 算法优于类中心算法和简单向量距离算法;特征比较明显,关键字比较集中的页面的查准率比较高,如色情类和反动言论类页面,而那些特征比较模糊、关键字分散的页面的查准率则比较低,如正常页面和非法交易页面。

5　结语

本文研究了自动分类技术在浏览器历史数据取证中的应用,实现了一个浏览器历史记录的自动分类取证系统,经实验证明该系统弥补了以往的浏览器历史数据取证工具的不足,提高了取证效率。此外,该系统在设计过程中利用了分层设计结构,使具有良好的扩展性,具体表现在以下方面:在浏览器数据挖掘层,可以很方便的增加对各种浏览器的支持;在数据挖掘层,可以很方便的采用其他分词算法、特征提取算法和分类算法。

参考文献:

[1] JONES KJ. Forensic analysis of internet explorer activity files[EB/OL]. http://www.foundstone.com/pdf/wp_index_dat.pdf, 2003-3-19.

[2] SALTON G, WONG A, YANG CS. A vector space model for automatic indexing[J]. Communication of the ACM, 1975, 18(11):613-620.

[3] 张国煊, 王小华, 周必水. 快速书面汉语自动分词系统及其算法设计[J]. 计算机研究与发展,1993,30(1): 61-65.

[4] SEBASTIANI F. Machine learning in automated text categorization [J]. ACM Computing Surveys (CSUR), 2002, 34(1): 1-47.

[5] YANG Y, PEDERSEN JO. A comparative study on feature selection in text categorization[A]. Proceedings of the Fourteenth International Conference on Machine Learning[C]. 1997.412-420.

[6] 高洁, 吉根林. 文本分类技术研究[J]. 计算机应用研究, 2004,(7): 28-30.

[7] DASARATHY BV. Nearest Neighbor (NN) norms: NN pattern classification techniques[M]. Los Alamitos, CA: IEEE CS Press, 1991.

(上接第 2426 页)

户偏爱的使用模式,是一种可以商用化的挖掘算法。

参考文献:

[1] 张峰, 常会友. Web 使用挖掘系统研制中的主要问题和对应策略[J]. 计算机科学, 2003, 30(6) :129-132.

[2] 邢东山, 沈钧毅. 从 Web 日志中挖掘用户浏览偏爱路径[J]. 计算机学报, 2003, 26(11):1518-1523.

[3] 战立强, 刘大昕. 基于访问路径树的 Web 频繁访问路径挖掘算法研究[J]. 计算机应用研究, 2005, 22(1):96-98.

[4] WANG L, MEINEL C. Behavior Recovery and Complicated Pattern Definition in Web Usage Mining[A]. Springer Verlag[C]. Berlin Heidelberg, 2004.531-543.

[5] BERENDT B, MOBASHER B. The Impact of Site Structure and User Environment on Session Reconstruction in Web Usage Analysis [A]. Proceedings of the WebKDD[C]. 2002.159-179.

[6] LI S, LING C. Mining the Most Interesting Web Access Associations [A]. WebNet World Conference on the WWW and Internet[C]. 2000.489-494.

[7] CHEN M-S, PARK JS. Data Mining for Path Traversal Patterns in a Web Environment[A]. Proceedings of the 16th International Conference on Distributed Computing Systems[C], 1996.