一种区分索引与信息的网页分类数学模型及证明

王树西[1] 夏增艳[2]

(对外经济贸易大学信息学院 北京 100029)[1] (北京邮电大学民族教育学院 北京 102209)[2]

摘 要 综述了网页分类的国内外研究现状,分析了网页分类的核心技术,包括思想、算法、公式、评测标准。为了打击非法网络传销,必须对传销网页进行准确识别与分类。根据网页中"最大内容段"的长度,计算出这个网页为"信息网页"的概率,通过严格的数学公式推导得到数学模型。本数学模型已经得到应用,"网络传销国家监控中心"用这个模型有效地对网络传销网页集合进行了分类。

关键词 网页分类,索引页面,信息页面,网络传销,网络传销国家监控中心

中图法分类号 TP392 **文献标识码** A

Classification Mathematical Model and Proof to Distinguish Index and Information Web Page

WANG Shu-xi[1] XIA Zeng-yan[2]

(Information Academy of the University of International Business and Economics, Beijing 100029, China)[1]

(National Institute of Education, Beijing University of Posts and Telecommunications, Beijing 102209, China)[2]

Abstract This paper surveyed domestic and international research of Web page classification, analyzed core technologies of Web page classification, including ideas, algorithms, formulas and evaluation criteria. In order to attack Internet Pyramid Selling, it is necessary to accurately identify and classify Internet Pyramid Selling Web pages. According to "maximum content length" of Web page, its "information page" probability is calculated. Web page classification mathematical model is deduced through strict formulas. Above mathematical model has been applied, and "National MLM Monitor Center" effectively classifies Internet Pyramid Selling Web pages using above model.

Keywords Web page classification, Indexed page information page, Internet pyramid selling, National MLM monitor center

1 引言

分类问题是一个很有意义的问题。将事物正确地进行分类,使杂乱无章的现实世界变得井井有条,这有助于人们正确地认识世界。互联网的迅猛发展和网页数量的剧增,使得人们对于网页分类的要求越来越迫切。简而言之,网页分类就是根据某一主题对网页集合进行自动分类,或者说,网页分类是使用机器学习的方法实现网页类别的自动标注。

网页分类已经在很多领域得到广泛应用,包括:信息检索、机器翻译、自动文摘、信息过滤、邮件分类等。例如,网页分类在搜索引擎中的用途包括:(1)根据不同的网页类型,制定相应的排序规则并进行相关性排序;(2)根据网页是索引页面还是信息页面,下载调度时做出不同的调度策略;(3)在网页信息抽取的时候,根据网页分类的结果制定不同的抽取策略;(4)在识别检索意图的时候,根据用户点击的网页类别推断检索意图。随着移动互联网进入4G时代,在大数据和云计算技术背景下研究网页分类,显得尤为重要。

本文综述国内外的网页分类研究现状;梳理网页分类的主要思路、核心算法、重要公式、评测机制;对大量网络传销网页进行统计分析,根据网络传销问题的实际需求,建立网页分类模型,来区分一个网络传销网页是索引页面还是信息页面;通过严格的数学公式推导,从理论上证明这个网页分类模型是正确的;最后指出这个网页分类模型的用途。

2 国内外研究现状

近年来,国内对网页分类的研究非常活跃,研究集中在两个方面:(1)提出新的网页分类算法;(2)研究网页分类算法的应用。

李晓黎、刘继敏、史忠植于2001年提出了一种将支持向量机与无监督聚类相结合的网页分类算法,给出了一种网页表示方法并应用于网页分类问题,该算法充分利用了SVM准确率高与无监督聚类速度快的优点。孙建涛、沈抖、陆玉昌、石纯一于2004年回顾了文本分类技术的研究状况,分析了网页的结构特征。侯翠琴、焦李成于2009年充分利用网页数据的超链接关系和文本信息,提出了一种用于网页分类的归纳式半监督学习算法:基于图的Co-training网页分类算法,并从理论上证明了算法的有效性。郑德权、张迪、赵铁军、于浩于2007年针对Blog网页的特点与规律,提出一种根据

本文受对外经济贸易大学"信息学院基金"(13YBLG02,X12511)资助。

王树西(1976—),男,博士,讲师,主要研究方向为商务智能等,E-mail:wangshuxi2006@sina.com;**夏增艳**(1979—),女,硕士,讲师,主要研究方向为数据分析。

网页结构和关键字计算相似度的方法来识别 Blog 网页，达到了较高的识别正确率。彭涛、左万利、赫枫龄、张长利于 2006 年通过对迭代产生的分类器进行优化组合，以及对网页结构的划分，寻找并利用网页集中蕴藏的规律综合计算特征权值，大大提高了网页分类的正确率和 F-measure 值。李宇峰、黄圣君、周志华于 2012 年提出了一种基于正则化的归纳式半监督多标记学习方法-MASS，在网页分类和基因功能分析问题上的实验结果验证了 MASS 方法的有效性。赵志滨、贾岩峰、姚兰、鲍玉斌于 2013 年针对含有丰富结构化数据的 Web 页面，提出了复用结构化数据抽取模板来进行 Web 页面主题识别的分类框架。单松巍、冯是聪、李晓明于 2003 年针对中文网页，比较研究了 CHI、IG、DF 以及 MI 特征选取方法。张茂元、邹春燕、卢正鼎于 2007 年为准确地分类网页，给出一种模糊网页分类的系统结构，给出一种通用学习规则，并提出一种变调整规则的单参数学习算法来加快参数学习速度。殷贤亮、李猛于 2007 年对互联网上大量存在的基于模板的网页，根据其半结构化的特点，提出了一种网页分块和主题信息自动提取算法。汤亚玲、崔志明于 2012 年提出一种结合用户行为特征分析的网页分类技术，该分类方法与多种统计学方法相结合实施网页分类均能有效地提高分类准确率。彭小刚、明仲、王海涛、周景洲于 2009 年研究了基于 wordNet 的类别可拓展网页分类系统，它使用关键词分类算法来提高分类准确率。王振宇、唐远华、郭力于 2012 年提出基于站点分层结构的网页分类与抽取，基于《知网》进行词语语义相似度计算。张婕、山岚于 2013 年研究了 CBC 算法在网页分类中的应用。左敬龙、余桂兰于 2011 年提出了将具有量子特性的 ACA 和 SVM 进行融合的中文网页分类方法。张青于 2014 年研究了移动互联网场景中客户特征分类技术。傅向华、刘国、陈冬剑于 2011 年针对 Web 页面分类方法一般只能处理小规模数据的问题，提出一种核心子集选择训练的大规模中文网页分类方法。宋军涛、周铜、杜庆灵于 2009 年提出了一种基于支持向量机和蚁群算法相结合的构造网页分类器的高效分类方法。陈沧于 2010 年研究了基于大规模类别体系的网页分类及在商品分类中的应用。孙聪凯于 2009 年研究了语义模型、近似推理算法及其在网页分类的应用。余桂兰、陈珂、左敬龙于 2014 年提出了一种基于云模型的并行蚁群-SVM 网页分类方法。秦兵、郑实福、刘挺、张刚、李生于 2002 年研究了可分性判据在中文网页分类中的应用。王天江、孔华武于 2007 年提出一种基于定性推理的网页分类方法。阎红灿、李敏强、任蕴丽、阎少宏于 2009 年针对 XML 网页特点，提出了一种结构和内容联合提取的 XML 网页分类研究方法。

国外对网页分类的研究也很活跃，近年来相关论文有很多。Qi Xiaoguang. Davison B D 于 2009 年总结了基于网页特征和分类算法的网页分类技术。Shen D. Yang Q，Chen Z 于 2007 年利用摘要对网页进行分类，它比纯文本方法要好。BroughtonV 于 2008 年通过同义词库的方法进行分类。Barzan Mozafari，Carlo Zaniolo 于 2009 年研究了基于简单贝叶斯模型而不失准确性的分类算法。Fiol-Roig G，Mir6-Julia M，Herraiz E 于 2011 年研究了数据挖掘技术在网页分类中的应用。Baykan EHenzinger MMarian L 等于 2011 年深入研究了基于 URL 主题分类的特征及算法。Sriurai W，Meesad P，Haruechmyasak C. 于 2010 年通过一个主题模型整合相邻页面，提高了网页分类性能。

国内外主要集中于研究网页分类的算法，从而提高网页分类的性能；而对网页分类模型的研究相对较少。

3 网页分类核心技术：思路、算法、公式、评测标准

网页分类的基本步骤是这样的：首先定义分类体系，将预先分类过的网页作为训练集，从训练集中得出分类模型（公式的相关参数），然后用训练得到的分类模型对其它网页进行分类。网页分类的一个关键问题，是特征词的选择及其权重分配。

网页分类算法很多，比较成熟的算法有：(1)朴素贝叶斯(Naive Bayes，NB)分类算法；(2)支持向量机(Support Vector Machine，SVM)分类算法；(3)K-近邻分类算法；(4)决策树分类算法；(5)神经网络(Neural network，NN)分类算法等。下面分别综述网页分类的算法、公式、评测标准。

3.1 朴素贝叶斯分类算法

贝叶斯分类(Naive Bayes，NB)的基础是：假设样本每个特征与其它特征都不相关。朴素贝叶斯分类器，是一种基于独立假设贝叶斯定理的简单概率分类器。朴素贝叶斯分类器的优势，在于根据少量的训练数据，估计出必要的参数。

概率模型分类器是一个条件概率模型：$p(C|F_1,\cdots,F_n)$。独立的类别变量 C 有若干类别，依赖于特征变量 $F_1,F_2,\cdots,F_n$。贝叶斯定理有以下公式：$p(C|F_1,\cdots,F_n)=\frac{1}{Z}p(C)\prod_{i=1}^{n}p(F_i|C)$，其中证据因子 Z 是一个只依赖于 $F_1,\cdots,F_n$ 的缩放因子，当特征变量的值已知时是一个常数。

3.2 支持向量机分类算法

支持向量机(Support Vector Machine，SVM)是由 Vapnik 于 1995 年提出的一种分类技术。所谓支持向量是指那些在间隔区边缘的训练样本点，“机”实际上是一个算法。SVM 方法提供了解决“维数灾难”问题的方法，SVM 的关键在于核函数。

支持向量机是一种二类分类模型，其基本模型定义为特征空间上间隔最大的线性分类器，即支持向量机的学习策略是间隔最大化，最终可转化为凸二次规划问题的求解。

线性可分的分类函数：

$f(x)=w^{\mathrm{T}}x+b$

线性不可分的分类函数：

$$f(x)=(\sum_{i=1}^{n}\alpha_i y_i x_i)^{\mathrm{T}}x+b$$

$$=\sum_{i=1}^{n}\alpha_i y_i\langle x_i,x\rangle+b$$

SVM 在文本分类领域有很多的应用。一个网页分类系统不仅是一个自然语言处理系统，也是一个典型的模式识别系统，系统的输入是需要进行分类处理的网页，系统的输出则是与网页关联的类别。

3.3 K-近邻分类算法

K-近邻(K-Nearest Neighbor，KNN)算法由 Cover 和 Hart 于 1968 年提出，理论上比较成熟，是最简单的机器学习算法之一，在网页分类中经常使用。

算法的思路是：样本 s 的特征空间中，有 K 个最相似(特征空间中最邻近)样本，如果这 K 个样本大多属于类别 C，则样本 s 也属于类别 C。K-近邻算法的核心，在于找到实例点

的邻居。这是因为，特征空间中两个实例点的距离反映出两个实例点之间相似性的程度。

两个实例点之间的距离，可以有多种表示方式：欧氏距离、曼哈顿距离、切比雪夫距离、闵可夫斯基距离(Minkowski Distance)、标准化欧氏距离(Standardized Euclidean distance)、马氏距离(Mahalanobis Distance)、巴氏距离(Bhattacharyya Distance)、汉明距离(Hamming distance)、夹角余弦(Cosine)、杰卡德相似系数(Jaccard Similarity Coefficient)、皮尔逊系数(Pearson Correlation Coefficient)。

K-近邻算法一个重要的问题是 K 值的定义，也就是选择多少个邻居，这对 K-近邻算法的结果会产生重大影响。

(1)如果选择较小的 K 值。意味着整体模型变得复杂，容易发生过拟合。

(2)如果选择较大的 K 值。意味着整体模型变得简单。

(3)$K=N$。错误，因为忽略了训练实例中大量有用信息。

(4)在实际应用中，K 值一般取一个比较小的数值。例如采用交叉验证法(简单来说，就是一部分样本做训练集，一部分做测试集)来选择最优 K 值。

算法步骤如下：

(1)对预料库文本预处理，形成文本向量 $V(v_1, v_2, \cdots, v_n)$。

(2)对测试文本进行分词、特征提取等，形成文本向量 $U(u_1, u_2, \cdots, u_n)$。

(3)计算待分类样本与训练样本间的相似度。

(4)对相似度由小到大排序，选取前 K 个文本。K 的取值没有定性的一个数值，需要随着实验的进行不断更改，选取到最合适的 K 值。

(5)在选取的 K 个文本中，分别统计待分类样本对于每个类别的次数，其计算公式为：

$$p(u,c_j)=\sum_{i=1}^{k}\mathrm{sim}(u,v)y(v_i,c_j)$$

其中，$y(v_i,c_j)$是类别属性函数，当 $v_i\in c_j$ 时，值为 1；当 $v_i\notin c_j$ 时，值为 0。

(6)分类决策函数为 $RC=\max c_j(p(v,c_j))$。

3.4 决策树分类算法

相比贝叶斯算法，决策树(Decision Tree)分类算法的优势在于：决策树构造过程不需要任何领域知识或参数设置。因此在实际应用中，对于探测式的知识发现，决策树更适用。

构造决策树的关键步骤是分裂属性，就是在某个节点处按照某一特征属性的不同，划分构造不同的分支，其目标是尽量让一个分裂子集中待分类项属于同一类别。分裂属性分为 3 种不同的情况：

(1)属性是离散值且不要求生成二叉决策树。此时用属性的每一个划分作为一个分支。

(2)属性是离散值且要求生成二叉决策树。此时使用属性划分的一个子集进行测试，按照“属于此子集”和“不属于此子集”分成两个分支。

(3)属性是连续值。

构造决策树的关键性内容是进行属性选择度量，属性选择度量是一种选择分裂准则，是将给定的类标记的训练集合的数据划分成个体类的启发式方法。属性选择度量算法有很多，一般使用自顶向下递归分治法，并采用不回溯的贪心策略。ID3 和 C4.5 就是其中两种常用算法。

决策树常用的公式为：

$$\Delta_i=I(parent)-\sum_{j=1}^{k_i}\frac{N(v_j)I(v_j)}{N}$$

其中，k_i 是属性 i 的值；N 是观测的数目；v_j 是根据属性 i 的第 j 个划分。

3.5 神经网络分类算法

神经网络(Neural network，NN)技术，也是常用的网页分类方法。神经网络是人工智能中的成熟技术，将神经网络用于网页分类，需要为每个分类建立一个神经网络，通过学习得到从输入单词(或者更复杂的特征词向量)到分类的非线性映射。其计算量和训练时间非常庞大。

3.6 相关公式

使用向量空间模型来衡量页面之间的相似度，计算公式如下：

$$\mathrm{Sim}(D_1,D_2)=\cos\theta=\frac{\sum_{k=1}^{N}w_{1k}\times w_{2k}}{\sqrt{\sum_{k=1}^{N}w_{1k}^2\times\sum_{k=1}^{N}w_{2k}^2}}$$

其中，W_{1k} 是第 k 个特征词在文本 D_1 中的权重。W_{2k} 是第 k 个特征词在文本 D_2 中的权重。

使用 TF-IDF 来考虑特征词项的权重，其计算公式如下：

$$TF\text{-}IDF(w)=TF(w)*IDF(w)$$

$$TF(w)=0.5+0.5\times nTerm/maxn$$

其中，$nTerm$ 是词 w 在给定文档中出现的次数，$maxn$ 是给定文档中词出现的最大次数。

$IDF(w)$是通过统计得到的词 w 的权重，$IDF(w)$的计算公式如下：$IDF(w)=\dfrac{\log_e(N/DF(w)+0.01)}{C}$。其中 N 是文档集合中的文档数，$DF(w)$是词 w 在文档集合中出现的文档数，C 为一个常数，其值为 $\log_e(N+0.01)$。

通过以上计算，可以得到网页与其链接页面的相似度值，对这些值进行排序，对于一个网页，可以得到按相似度排序的链接页面集合。

在进行特征提取时一般采用开方检验特征提取法，计算特征项 t 和类别 c 的相关性公式：

$$\chi^2(t,c)=\frac{N\times(AD\text{-}CB)^2}{(A+C)\times(B+D)\times(A+B)\times(C+D)}$$

分类时采用 SVM 分类器，核函数采用径向基函数：

$$f(x)=sign(\sum_{i=1}^{l}\alpha_i K_\gamma(|x-x_i|)-b)$$

最通用的判定规则是采用高斯函数：

$$K_\gamma(|x-x_i|)=\exp\{-\frac{|x-x_i|^2}{\sigma^2}\}$$

3.7 评测标准

在网页分类系统中，查全率和查准率反映了分类质量的两个不同方面，两者必须综合考虑，表示为 F1 测试值。下面是网页分类常用的评测公式：

$$F_1=\frac{2pr}{p+r}$$

其中，p 为查准率，r 为查全率，m 为训练集类别数。由于 Macro-F1 可以显示分类器的整体性能，因此，很多文献使用 Macro-F1 值作为分类器性能的评价标准。

$$Macro-F_1=\frac{1}{m}\sum_{i=0}^{m=1}F_{1i}$$

在网页分类系统的评测过程中，一般首先计算每个类的 F_1 值，然后计算 $Macro-F_1$ 的值。

4 一个打击网络传销的网页分类数学模型

近几年来，网络传销甚嚣尘上，而且伪装性、欺骗性越来越强，很多人上当受骗，带来很多人间悲剧。网络传销已经成为社会公害，造成严重的社会后果，成为社会大患。必须整顿网络秩序，对网络传销进行严厉打击，首先要做的就是要对网络传销网页进行准确识别与分类。

从 2012 年 6 月开始，国家工商行政管理总局（以下简称“总局”）组建了“网络传销国家监控中心”，并委托对外经济贸易大学电子商务研究所进行建设。经过 2 年多的辛苦建设，“网络传销国家监控中心”已经具有相当的规模，获得一系列重要成果，抓获了大批非法网络传销网站。

4.1 网络传销网页特点分析

网络传销越来越隐蔽，网络传销网页也越来越难以识别。一般来说，网络传销网页主要包括如下几种：

（1）赤裸裸宣传网络传销。

这种网页打着“网络投资”、“网络广告”、“网络赚钱”的幌子，大肆宣称其非法网络传销，并提供银行账号让受骗者向里面汇款。随着工商、公安部门的联合打击，这类赤裸裸宣传网络传销的网站（网页）越来越少。

（2）宣传“电子商务新理念”。

随着电子商务的兴起，很多非法网络传销披上了“电子商务”的华丽外衣，宣传所谓的“电子商务新理念”，也就是加入之后有回扣，发展下线越多回扣越多。在这个幌子下面，许多人陷入了“网络购物”的泥淖。这样的网站（网页）现在有很多。

（3）“高科技公司发行原始股份”

很多非法网络传销披上了“高科技”的外衣，在网站（网页）上大肆宣称“高科技发行原始股份”，给不明真相的人们种种虚假的承诺。在“高科技”炫目的光环下，在“原始股份”高额回报的诱惑下，很多人面对这样的网站（网页）宣传无法自持，掉进了陷阱。

（4）诱导私下“网络聊天”

这种网站（网页）“简明扼要”地说明非法网络传销的种种“好处”，然后留下电子邮箱或者即时聊天工具的账号（如 qq、skype 等），诱导受害者进行私下网络聊天。这种网站（网页）的宣传内容往往非常短小，欺骗工作主要在私下网络聊天的时候进行。

（5）加密的网站

这类网络传销网站（网页）只有简单的 2 个文本框：用户名文本框和密码文本框，除此之外，没有任何宣传网络传销的文字、图片内容。这类网络传销网站的用户，是那些已经被彻底洗脑、思想被严格控制的网络传销“铁杆”用户，他们已经交了钱，拥有进入这些网络传销网站的用户名和密码，进入网站之后进行网络传销行为。

4.2 网络传销网页特征统计

从事非法网络传销行为的网站（网页）有多种不同的类型。但是可以将非法网络传销网页分为两类：索引网页；信息网页。

“索引网页”比较简短，宣传网络传销的内容较少，主要是诱导客户通过即时聊天工具进行私下宣传，或者链接到其他网络传销页面。“信息网页”通过大量的内容（文本、图像、图表）宣传网络传销。

对现有的大量网络传销网页进行统计后，发现一个规律：如果一个网页中的“最大内容段”长度大于 400 个汉字，则基本可以断定该网页是“信息网页”，否则就是“索引网页”。

通过上述统计规律，可以对大量网络传销网页进行分类：“索引网页”、“信息网页”。

但这种根据“最大内容段”进行网页分类的方法过于简单。我们希望将网页中“最大内容段”的长度映射为一个概率值，根据网页中“最大内容段”的长短，得到这个网页为“信息网页”的概率，也就是得到一个公式模型。针对上述问题，通过公式推导得出一个网页为“信息网页”的公式模型。

4.3 数学模型推导

（a）条件的自然语言描述：

网页中“最大内容段”的长度越大，其所对应的“信息网页”概率也越大，最大为 1。数据统计结果表明，当“最大内容段”的长度为 400 个汉字的时候，这个网页是“信息网页”的概率为 0.7。“最大内容段”字数为 0 的时候，该页成为“信息网页”的概率为 0。

（b）条件的形式化描述：

设映射 $y=f(x)$ 满足下面的特性，构造函数 $y=f(x)$。自变量 x 为“最大内容段”的字数，因变量 y 为这个网页是“信息网页”的概率。

（1）定义域：$[0,\infty)$；

（2）值域：$[0,1]$；

（3）$y=f(x)$ 是单调递增的；

（4）$x=400$ 时，$y=0.7$；

（5）$x=0$ 时，$y=0$。

（c）公式模型推导过程：

$1-f(x)=cf'(x)$

可以写成：$1-y=c\frac{dy}{dx}$

也就是：$dx=c\frac{dy}{1-y}$

公式两边积分，得到：$x=-c\ln(1-y)+C$

或者写成：$\ln(1-y)=-\frac{x}{c}+\frac{C}{c}$

上述公式可以写成：$1-y=e^{(-\frac{x}{c}+\frac{C}{c})}$

也就是：$y=1-e^{(-\frac{x}{c}+\frac{C}{c})}$

根据边界条件（$x=0$ 时，$y=0$），得到：$C=0$。

取 $c=1/k$，得到：$y=1-e^{-kx}$。

根据已知条件：$x=400$ 时，$y=0.7$，得到：$k=0.003$。

也就是：$y=1-e^{-0.003x}$

4.4 数学模型分析

上述数学模型（$y=1-e^{-0.003x}$）可以更加直观地表述为图 1 所示。

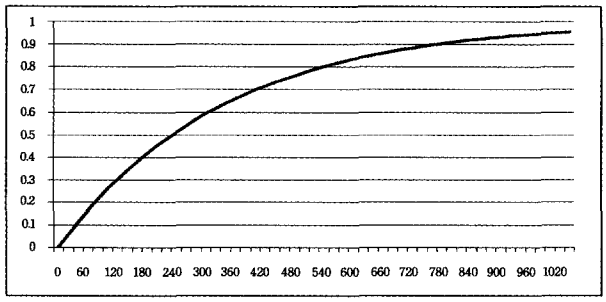

图1 网络传销网页为"信息网页"的概率

在上述图形中，横轴为网络传销网页中"最大内容段"的长度，纵轴为这个网络传销网页为"信息网页"的概率。从上述图形可以看出，对于一个非法的网络传销网页，"最大内容段"长度越大，那么这个网页为"信息网页"的概率越大。

根据上述网页分类的数学模型($y=1-e^{-0.003x}$)，当 $x=235$ 时，$y=0.505891426$。也就是说，网络传销网页中"最大内容段"长度为 235 时，这个网页为"信息网页"的概率为 0.505891426。也就是说，如果网络传销网页中"最大内容段"长度大于 235，那么这个网页可以分类为"信息网页"，否则分类为"索引网页"。

4.5 数学模型应用

随着互联网的迅猛发展，网页的数量也在快速增长。很多网页属于"索引型"网页，里面没有具体内容，只是一些目录链接；更多的网页属于"信息性"网页，里面是某一专题的内容。

根据上述数学模型，"网络传销国家监控中心"已经对网络传销网页进行有效的分类，并取得了良好效果。根据上述数学模型，可以按照"索引型"与"信息型"这个"二值分类"标准对网页进行分类。必须指出，上述数学模型的参数是针对网络传销网页的，对于不同领域的网页进行分类，必须得到相应的参数。

结束语 因特网上的网页数量急剧膨胀，如何从海量的网页中迅速、准确地搜索用户感兴趣的信息是对网页分类技术的挑战。针对传销网站的准确识别与分类问题，对网页分类技术进行了研究和探讨，研究工作具有较大的实际意义。

网页分类的过程，可以看作是网页集合与类别集合的映射。本文综述了国内外网页分类研究现状，分析了网页分类的核心技术，包括思想、算法、公式、评测标准。

为了打击非法网络传销，"网络传销国家监控中心"针对抓获的大量网络传销网页进行统计分析，根据网页中"最大内容段"的长短，得到这个网页为"信息网页"的概率，通过严格的数学公式推导得到数学模型。根据本数学模型，"网络传销国家监控中心"已经有效地对传销网页集合进行分类。

下一步的工作，将进一步加深对国内外的研究现状的了解，尤其是对国内外近两年内有关网页分类技术方面的研究进行深入了解。论文的核心是通过网页中"最大内容段"的长度大小来推断该网页是"信息网页"的概率。简单地将计算概率大小的公式称为是网页分类数据模型有些牵强，缺乏依据，下一步工作中，将寻找相关依据。

下一步的工作，将在本数学模型的基础上进行改进，根据"网络传销国家监控中心"实际工作的需要，推导出"多维分类"标准下的网页分类数学模型，加大创新力度。

随着大数据、云计算、物联网、移动通信的迅速发展，网页分类的研究工作必须与时俱进。特别是移动通信已经进入4G时代，微博、微信的数量迅猛增长，对微博、微信进行分类的工作正在迅速展开。此外，网络社区的研究工作如火如荼，将网页分类的研究工作与网络社区的研究工作结合在一起，将是很有意思的事情。

参 考 文 献

[1] Qi Xiao-guang, Davison B D. Web Page Classification: Features and AtgorithmsI[J]. ACM Computing Surveys(CSUR), 2009, 41(2): 12-42

[2] Shen D, Yang Q, Chen Z. Noise Reduction Through Summarization for Web Page classificalion[J]. Information Processing & Management, 2007, 43(6): 1735-1747

[3] Broughton V. A faceted classification as the basis of a faceted terminology: Conversion of a classified structure to thesaurus format in the Bliss Bibliographic Classification (2nd Ed.)[J]. Axiomathes, 2008, 18(2): 193-210

[4] Mozafari B, Zaniolo C. Publishing naive bayesian classifiers: Privacy without accuracy loss[C]// Proc of the VLDB Endowment. New York: ACM, 2009: 1173-1185

[5] Fiol-Roig G, Mir6-Julia M, Herraiz E. Data mining techniques for Web page classification[J]. Highlights in Practical Applications of Agents and Muhiagent Systems, 2011, 89: 61-68

[6] Baykan E, Henzinger M, Marian L, et al. A comprehensive study of features and algorithms for URL-based topic classification [J]. ACM Transactions on the Web(TWEB), 2011, 5(3): 15

[7] Sriurai W, Meesad P, Haruechmyasak C. Improving Web page classification by integrating neighboring pages via a topic model [C]// Proceedings of IICS2010. 2010: 238-246

[8] 李晓黎，刘继敏，史忠植. 基于支持向量机与无监督聚类相结合的中文网页分类器[J]. 计算机学报，2001(9)

[9] 侯翠琴，焦李成. 基于图的 Co-Training 网页分类[J]. 电子学报，2009(10)

[10] 鲁明羽，沈抖，郭崇慧，等. 面向网页分类的网页摘要方法[J]. 电子学报，2006(8)

[11] 郑德权，张迪，赵铁军，等. Blog 网页分类与识别技术研究[J]. 通信学报，2007(12)

[12] 孙建涛，沈抖，陆玉昌，等. 网页分类技术[J]. 清华大学学报：自然科学版，2004(1)

[13] 彭涛，左万利，赫枫龄，等. 基于粒子群优化算法的网页分类技术[J]. 计算机研究与发展，2006(3)

[14] 李宇峰，黄圣君，周志华. 一种基于正则化的半监督多标记学习方法[J]. 计算机研究与发展，2012(6)

[15] 赵志滨，贾岩峰，姚兰，等. 含有丰富结构化数据的 Web 页面分类技术的研究[J]. 计算机研究与发展，2013(1)

[16] 单松巍，冯是聪，李晓明. 几种典型特征选取方法在中文网页分类上的效果比较[J]. 计算机工程与应用，2003(22)

[17] 张茂元，邹春燕，卢正鼎. 一种基于变调整学习规则的模糊网页分类方法研究[J]. 计算机研究与发展，2007(1)

[18] 殷贤亮，李猛. 基于分块的网页主题信息自动提取算法[J]. 华中科技大学学报：自然科学版，2007(10)

[19] 汤亚玲，崔志明. 行为特征分析模式下的网页分类技术研究[J]. 计算机工程，2012(20)

[20] 彭小刚，明仲，王海涛，等. 基于 wordNet 的类别可拓展网页分类系统[J]. 深圳大学学报：理工版，2009(2)

[21] 王振宇，唐远华，郭力. 面向分层结构的网页分类与抓取[J]. 计算机工程与科学，2012(11)

[22] 张婕,山岚. CBC 算法在网页分类中的应用研究[J]. 北京化工大学学报:自然科学版,2013(1)
[23] 左敬龙,余桂兰. 具有量子特性的 ACA-SVM 网页分类方法[J]. 计算机工程与应用,2011(12)
[24] 张青. 移动互联网场景中客户特征分类技术研究[J]. 电信科学,2014(1)
[25] 傅向华,刘国,陈冬剑. 一种核心子集选择训练的大规模中文网页分类方法[J]. 小型微型计算机系统,2011(8)
[26] 宋军涛,周铜,杜庆灵. 支持向量机和蚁群算法的网页分类研究[J]. 计算机工程与应用,2009(17)
[27] 陈沧. 基于大规模类别体系的网页分类及在商品分类中的应用研究[D]. 扬州:扬州大学,2010
[28] 孙聪凯. 语义模型、近似推理算法及其在网页分类的应用[D]. 上海:上海交通大学,2009
[29] 余桂兰,陈珂,左敬龙. 基于云模型的并行蚁群-SVM 分类方法[J]. 计算机技术与发展,2014(4)
[30] 秦兵,郑实福,刘挺,等. 可分性判据在中文网页分类中的应用[J]. 微处理机,2002(1)
[31] 王天江,孔华武. 一种基于定性推理的网页分类方法[J]. 计算机工程与应用,2007(9)
[32] 阎红灿,李敏强,任蕴丽,等. 结构和内容联合提取的 XML 网页分类研究[J]. 天津大学学报:社会科学版,2009(3)

(上接第 287 页)

设施作为服务,提供商负责 Hypervisor 层以下层次的安全责任,即只负责物理安全、环境安全和虚拟化安全等这些安全控制,而用户负责操作系统安全、应用安全和数据安全。SaaS 云提供整个服务,提供商不仅负责物理、环境和虚拟化安全,还必须负责操作系统、应用和数据的安全控制。

按需防护的安全框架是高效的。给所有服务使用强安全策略降低了云服务的速度。根据用户使用的服务类型、用户指定的安全要求以及接入网络特点等区别使用不同的安全保护措施,有助于提高云服务的整体速度。

按需防护的安全框架的另一个特点是简单。用户输入层的 3 个参数都可以设置默认值,用户选择某个服务,系统便可自动判断安全等级、服务类型和接入网络的风险等级,进而提供相应的安全防护。用户也可以更改安全等级。

结束语 云计算是当前发展十分迅速的新兴产业,具有广阔的发展前景,但同时其所面临的安全技术挑战也是前所未有的。本文综合分析了云计算的安全目标、各类安全风险及防范对策,提出了按需防护的安全框架,该框架的用户输入层共有 3 个参数:安全等级、服务类型和接入网络的风险等级。每个云服务都有本身的默认安全等级,用户可以根据实际情况进行修改,也可以不修改。服务是用户申请使用的云服务的类型,系统可以自动获取。接入网络的风险等级可以由系统根据终端位置和 IP 自动判断。所以,在实际使用过程中,用户只需要设置 0－1 个参数,操作非常便利。云计算安全并不仅仅是技术问题,它还涉及标准化、监管模式、法律法规等诸多方面。因此,仅从技术角度出发探索解决云计算安全问题是不够的,需要信息安全学术界、产业界以及政府相关部门的共同努力才能实现[19]。

参 考 文 献

[1] 俞能海,郝卓,徐甲甲,等. 云安全研究进展综述[J]. 电子学报,2013,41(5):371-381
[2] Chen Z G, Liu L P, Liu A F. Trust-sensitive Web service composition strategy based on black and white board[J]. Journal on Communications, 2010, 31(6): 25-35
[3] 邓谦. 基于 Hadoop 的云计算安全机制研究[D]. 南京:南京邮电大学,2013
[4] 杨凯. 银联数据异地灾难备份架构设计探讨[J]. 中国金融电脑,2005,9(9):51-54
[5] Damiani E, Vimercati D C, Paraboschi S. A reputation based approach for choosing reliable resources in peer-to-peer net-works [C]// Proceedings of the 9th ACM Conference on Computer and Communications Security. 2002: 18-22
[6] Jurca R, Faltingsi B. Eliciting truthful feed-back for binary reputation mechanisms [C]// Proceedings of the 2004 IEEE/WIC/ACM International Conference on Web Intelligence. 2004: 214-220
[7] 陈志刚,刘莉平,刘安丰. 基于黑白板的信任敏感服务组合策略[J]. 通信学报,2010,31(6):25-35
[8] 潘静,徐锋,吕建. 面向可信服务选取的基于声誉的推荐者发现方法[J]. 软件学报,2010,21(2):388-400
[9] 胡春华,刘济波,刘建勋. 云计算环境下基于信任深化及集合的服务选择[J]. 通信学报,2011,32(7):71-79
[10] 于洋洋,虞慧群,范贵生. 一种云存储数据完整性验证方法[J]. 华东理工大学学报,2013,39(4):211-216
[11] 颜湘涛,李益发. 基于消息认证函数的云端数据完整性检测方案[J]. 电子与信息学报,2013,35(2):310-313
[12] 安玉,蒋天发,吴有林. 一种基于量子保密通信及信息隐藏协议方案[J]. 武汉大学学报,2012,45(3):394-398
[13] 李顺东,王道顺. 基于同态加密的高效多方保密计算[J]. 电子学报,2013,41(4):798-803
[14] Pan J, Xu F, Lv J. Reputation-based recommender discovery approach for service selection[J]. Chinese Journal of Software, 2010, 21(2): 388-400
[15] Ryan M D. Cloud computing security: The scientific challenge, and a survey of solutions[J]. The Journal of Systems and Software, 2013, 86(5): 2263-2268
[16] Van-Hau P, Dacier M. Honeypot Trace Forensics: The Observation Viewpoint Matters[J]. Future Generation Computer System, 2011, 27(5): 539-546
[17] Shpantzer G. Implementing Hardware Roots of Trust: The Trusted Platform Module Comes of Age[J]. SANS Analyst Program, 2013, 40(6): 1-15
[18] Liu H. A new form of DOS attack in a cloud and its avoidance mechanism[C]// Proceedings of the 2010 ACM Work-shop on Cloud Computing Security Workshop. New York, USA: ACM Press, 2010
[19] 冯登国,张敏,张妍,等. 云计算安全研究[J]. 软件学报,2011,22(1):71-83
[20] CSA. Security guidance for critical areas of focus in cloud computing v3.0[OL]. https://cloudsecurityalliance.org/guidance/csaguide.v3.0.pdf
[21] Thomas C. OW2 and the Open Cloud Industry Ecosystem[OL]. www.ciecloud.org/2013