

基于网络社团结构的 Web 内容分级算法及其性能分析

刘 艳^{1,2} 王 泰¹

(华中师范大学国家数字化学习工程技术研究中心 武汉 430079)¹

(武汉理工大学教育技术学系 武汉 430079)²

摘 要 万维网内容因其海量性、形式多样性和缺乏语义描述等特征,给内容分级实时自动化处理带来了巨大挑战。本算法充分利用相近主题网页聚合成内容社团的万维网结构特性,在对请求内容分级网页进行处理的同时,通过网络社团检测方法来自动获取其他更多相近内容的网页,以提升网页内容分级处理效率;此外,它能很好地融入现有网络内容第三方分级系统。理论分析证明,本算法能显著提升万维网内容分级处理效率。

关键词 网络内容分级,网络社团结构,第三方分级

中图法分类号 TP311.5 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2015.3.042

Web Content Rating Algorithm Based on Network Community Structure and its Performance Analysis

LIU Yan^{1,2} WANG Tai¹

(National Engineering Research Center for E-learning, Huazhong Normal University, Wuhan 430079, China)¹

(Department of Educational Technology, Wuhan University of Technology, Wuhan 430079, China)²

Abstract The Web contents are massive, diverse and semantics-missing, which bring significant challenges to the content-rating. This new algorithm takes full advantage of the Web structure feature that the similar topic webpages aggregate into the Web community, and uses the Web community detection algorithm to rate more content-similar webpages automatically when rating one webpage. In addition, it can be used in the current third-party content-rating system. Theory analysis shows that this algorithm significantly raises the efficiency of Web contents rating.

Keywords Web content rating, Network community structure, Third-party rating

1 引言

万维网因其信息丰富、应用多样和操作方便,已成为最常用的网络内容发布和获取媒介。随着万维网相关技术和应用服务的不断升级,万维网上的内容类型和传播方式还在不断扩展,信息发布和获取也日益便捷。但这种发展也使不良网络内容传播和获取更加方便,从而严重影响青少年健康成长、社会稳定和国家安全。因此,万维网上的内容分级和管理已成为国家、社会和研究机构关注的热点。

从 90 年代中期开始,许多研究机构和公司就陆续提出了各种网络内容分级解决方案。W3C 的 PICS 工作组曾最早于 1996 年提出过第三方分级技术^[1],但该技术只针对静态网页形式,且原则上要求分析所有网页上的内容,不能适应当前万维网海量数据、动态生成和快速更新的实际情况。此后,互联网内容分级协会(ICRA)又提出了基于自分级技术的内容过滤方案^[2,3],但该解决方案要求内容发布者在发布内容的同时提交准确描述其内容的语义标签,这种自律性要求在现实环境中难以确保。万维网内容的海量性使得人工分级不具有可实现性,为此,Mark Hepple 等人提出了基于自然语言处理的网页内容分级和过滤技术^[4],该技术仅局限在对文字内容

进行分析,无法对视频、游戏等内容自动分析,且它仅支持少数几种欧洲国家语言。另外,我国的“绿色上网标准特设项目组”于 2006 年提出了基于 PC 终端的互联网内容过滤技术^[5,6],但该技术方案依赖 PC 终端进行内容实时分析,准确性、实时性和灵活性不佳,且该方法的实现依赖于网络用户自律性,这在现实环境中也难以满足。

迄今为止,还没有任何一种技术能完全解决万维网内容分级和过滤问题,这是由于万维网的下述特性使得对网络内容分级的高效处理非常困难。

(1)海量内容:Google 于 2009 年就宣布已完成 1000 亿个 Web 网页的检索^[7],并且中国互联网络信息中心于 2013 年 6 月发布的第 89 期《互联网发展信息与动态》^[7]显示,万维网服务器数量从 2009 年的 2 亿台左右已增加到 2013 年的 6 亿多,增长约 3 倍^[8],从其增长率不难推断,目前 Web 上的网页总量非常巨大。

(2)缺乏语义信息:现有的 Web 架构是面向人而非机器的,其基本语言 HTML 只负责描述如何呈现内容,不包含有关内容的语义信息,这使得通过机器来实现对网络内容的自动理解、推理和分级变得非常困难^[9]。

(3)内容形态多样:万维网内容由包括文字、动画、图像、

到稿日期:2014-04-02 返修日期:2014-07-31

刘 艳(1973—),女,博士,副教授,主要研究方向为网络内容分级技术标准、教育技术应用,E-mail:ly_zzx@163.com;王 泰(1981—),男,博士,副教授,主要研究方向为网络内容分级、多媒体内容分发、无线网状网。

视频、音频、游戏等多种形态的内容随意混合在一起,准确定位不同形态内容的区域,并综合应用各种智能分析技术来确定内容级别仍然是一个非常困难的问题。

(4)内容易变:许多网站上的内容在不断变化,新的网页不断生成,旧的网页被删除,同一个网页上的内容也在不断变化,实时跟踪这些内容的变化并及时更新相关网页的内容标签也非常困难。

在深入分析本领域已有研究工作和万维网拓扑结构特点的基础上,本文提出了一种网络内容分级新方法。其主要特点是,充分利用万维网社团结构特征^[10],在对网页进行分级处理的同时,把那些已完成内容分级的网页作为“种子”网页,调用万维网社团结构检测算法来搜寻与“种子”网页同处一个社团的其他网页,并依据万维网社团组成网页具有相似内容主题的性质,给这些网页打上相似的内容分级标签。理论分析证明,它能显著提升网络内容分级处理的效率。

2 基于网络社团结构的万维网内容分级技术机制分析

万维网可看成由海量网页构成的一个巨大的复杂网络,其节点就是各个网页,而连接各个节点的边则是存在于网页之间的超链接。对万维网拓扑结构的研究表明,像万维网这类复杂网络的链接分布并非完全随机的,而是表现出一种明显的社团结构性^[11];整个网络由若干个大大小小的社团构成,每个社团内部的节点之间的连接相对较为紧密,各个社团之间的连接相对比较稀疏。同时,万维网社区结构又是分层次的,即某一层次的社团结构又可进一步细分成更小的若干社团^[12],如图 1 所示。

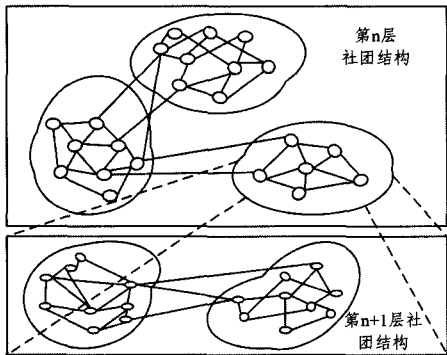


图 1 万维网社团结构图

对于万维网而言,其社团结构的内在形成机制是:网络内容发布者为提高其内容被用户浏览到的概率,会尽量使其内容与其他已发布的具有相似主题的网页通过超链接关联起来。因此,当我们浏览某一主题网页时,常常会发现连接到其他相同主题网页的超链接,对其他主题的网页也是如此,这充分显示了 Web 网页按内容主题聚合成社团的特性。另外,我们在选择网页上某具体内容项时,会发现新打开的网页除了对该内容的详细描述外,还包含其他对同一内容进行描述的网页链接,这充分显示了万维网社团的层次特性。

上述万维网分层社团结构为我们进一步提升网络内容分级效率提供了新的思路。

因为在万维网中,相似内容主题的网页聚合成紧密关联的社团,所以在完成对某一网页的内容进行分析和分级后,可以把该网页作为“种子”网页,利用网络社团检测方法来搜寻

处于同一社团内的其他网页,并赋予相似的内容分级标签。即根据社团结构背后所反映的“物以类聚”的原理,我们认为,与不良内容网页同处一社团的其他网页在很大概率上也会包含不良内容,而与正常内容网页同处一社团的其他网页在很大概率上包含的是正常内容。

如上所述,万维网社团结构的分层特性对应于网络内容主题的细分。具体来说,上层的社团结构对应于较一般的内容描述,比如“非普遍适宜内容”,而越下层的社团结构则对应于更具体的内容描述,比如“血腥的新闻图片”(允许成人浏览,而少儿则不宜)。这样当我们确定了某一层次的万维网社团所对应的内容描述不适合某一具体人群时,则该网络社团内的所有下级层次网络社团都不适合该人群。因此,当对某一未创建内容标签的网页进行分级时,若其上级某层社团结构已被描述为不良内容,则可直接过滤,而无须再进行内容分析。

基于上述原理,本文提出一种新的第三方分级算法。该算法充分利用万维网社团结构所提供的网络内容关联特性,结合网络社团结构检测和语义标签技术,以提升现有网络内容第三方分级处理的效率。当然,考虑到可能存在误判,在实际产品开发中,可通过集成不良内容举报机制来及时纠正内容误判。

3 基于网络社团结构的万维网内容分级算法

根据上述基于万维网社团结构的内容分级技术机制,我们设计得到如图 2 所示的万维网内容分级处理算法框架。

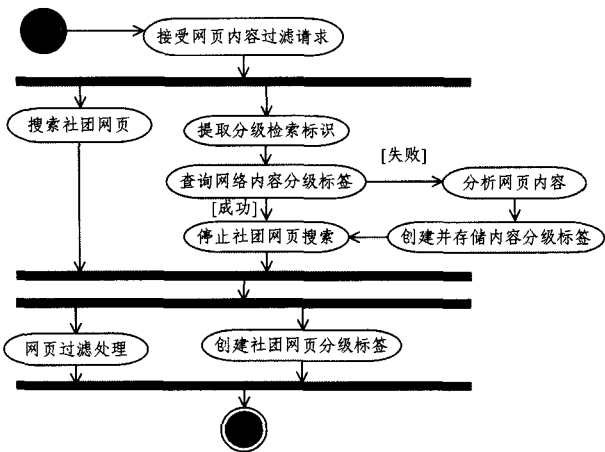


图 2 基于网络社团结构的万维网内容技术原理

该算法具体描述如下:

输入:待过滤网页地址 url,网页访问用户的过滤规则 rules,标签库 LabelBase
输出:网页过滤指示 block
Procedure WebContentRating(url, rules, LabelBase)
begin
 Do step (1) and (2) par-do
 (1) i←0; //待过滤网页所处社团网页检测
 while stop=FALSE do
 URLs[i]←EAGLE(url) /* 采用网络社团结构检测算法
 i←i+1
 EAGLE^[13]抓取“种子”网页所处
 社团中的其它网页 */
 end while
 (2) /* 网页过滤处理 */
 (2.1) id←ExtractID(url) //提取待过滤网页标签库查询标识

```

(2.2) label←Query(LabelBase,id) //标签库查询
(2.3) if lable=NULL
    rating ← Analyse(url) //通过网页内容分析得到其分级描述
    label←CreateLabel(rating) /* 依据网络内容分级标准,创建网
        页内容分级标签 */
    Add(LabelBase,id,label) //把新生成的标签添加到标签库中
then
    stop ← TRUE
end if
Do step (3) and (4) par-do
(3) /* 依据分级标签给出过滤建议 */
    block ← GetFilteringDesition(label,rules)
(4)/* 赋予所有检测到的同社区网页与待检测网页相同的标签值,
    并添加到标签库中 */
    for j=0 to i-1 do
        id ← ExtractID(URLs[i])
        Add(LabelBase,id,label)
    end for
end

```

由上可知,本算法具有下述特点:利用万维网的社团结构拓扑特性,在对当前网页进行分级处理的同时,充分获取与当前网页具有相似内容特性的其他网页,并给予相似的语义标签,从而极大地提高了网页分级处理效率。本算法是在现有第三方分级技术的基本流程中融入万维网社团结构检测过程,因此,该处理过程可以将独立的产品形态整合到已开发的第三方分级系统中,从而在有效保护前期研发投资的基础上,增强现有系统性能。

4 算法性能分析

为分析并评估本文所提出的万维网内容分级处理算法的性能,我们给出了本算法的数学模型,并将其与现有的、未考虑社团结构内容关联性的第三方分级处理算法进行对比。

值得说明的是,无论是本算法,还是现有的第三方分级处理算法,在实现过程中都可采用基于人工和基于内容智能分析这两种内容分析方式,也都可结合语义标签技术。因此,为使得性能比较更加客观,我们将假定这两种技术采用同样的网络内容分析方式,并且都使用语义标签技术。

4.1 符号说明

在后续性能分析中,我们将用到下述符号:

1. 设 N 为 Web 网页总数。

2. 设语义标签查询时长为 t_q 个单位时间,网页内容分析平均时长为 t_p 个单位时间。又因为内容分析算法的计算复杂度远高于数据库查询,所以语义标签查询时长 t_q 远小于网页内容分析平均时长 t_p ,在下述算法性能比较计算中,总将 t_q 近似为 0。

3. 设在单位时间内,由单个“种子”网页出发进行网络社团检测时所获得新网页数为 $x(x > 0, x \in R)$ 。

4.2 理论分析

1. 与 Web 社团结构无关的几个通用公式

若可查询到当前待处理网页的语义标签,考虑到标签查询只是一个数据库查询操作,其处理时间相对于网页内容分析时长可忽略不计,因此,可马上启动新的网页处理。

设 n_1 为过去已完成内容分级处理的网页数,则待处理网页存在语义标签的概率 p 为:

$$p = \frac{n_1}{N} \quad (1)$$

设 X 表示存在语义标签的待处理网页连续出现次数的随机变量,则其概率分布为:

$$P(X=k) = \left(\frac{n_1}{N}\right)^k * (1 - \frac{n_1}{N}), k=0,1,2,\dots \quad (2)$$

因此,此次分级处理的网页数为:

$$n_2 = 1 + E(X) = 1 + \frac{N}{N - n_1} \quad (3)$$

2. 利用了 Web 社团结构特点的其他公式

设在某一时刻已检测出的网页数为 n (不包括当前准备用于启动检测的“种子”网页);又设由当前“种子”网页将启动 m 次检测,以得到 m 个不同于当前“种子”网页的其它网页。

因为在新的检测过程中检测到的新网页有可能已包含在先前已检测出的网页中,所以当前检测所实际新增网页数有可能小于 m 。基本关系如下所示。

设随机变量 $A_i (1 \leq i \leq m, i \in N)$ 表示第 i 次检测出的网页与先前已检测出网页的关系:

$$A_i = \begin{cases} 1, & \text{已包含于已检测网页中} \\ 0, & \text{未包含于已检测网页中} \end{cases}$$

简单推导得:

$$P(A_i = 1) \approx \frac{n}{N-1} \quad (4)$$

$$E(A_i) = \frac{n}{N-1} \quad (5)$$

设随机变量 $B = \sum_{i=1}^m A_i$ 表示当前 m 次检测中所重复的网页数。

$$E(B) = E(\sum_{i=1}^m A_i) = \sum_{i=1}^m E(A_i) = \frac{m * n}{N-1} \quad (6)$$

所以,当前 m 次检测所新增的实际网页数 z 为:

$$z = m - \frac{m * n}{N-1} \quad (7)$$

3. 若 y_n 为准备分级处理第 n 个网页时已检测到的网页数(同时也是“种子”网页数),由式(7)可推得:

$$(i) n=1 \text{ 时}, y_1=0; \quad (8)$$

$$(ii) n=2 \text{ 时}, y_2=1; \quad (9)$$

$$(iii) n > 2 \text{ 时},$$

$$y_n = (1 - \frac{t_p * x}{N-1})^{n-2} + (1 + t_p * x) * (\sum_{i=0}^{n-3} (1 - \frac{t_p * x}{N-1})^i) \quad (10)$$

4. 相关参数设定和性能比较

Google 于 2008 年宣布已完成 1000 亿个 Web 网页的检索^[7],考虑到 Web 还在快速膨胀,在下述性能评估中,我们将 Web 网页数 N 近似为 $+\infty$ 。

以下对第三方分级处理传统算法和本算法分别进行计算。

(1) 未考虑社团结构内容关联性的第三方分级处理算法

设 z_n 表示采用现有第三方分级算法完成第 n 次分级处理时,扣除重复处理网页后的实际检测网页数。

由式(3)得到检测完成网页数:

$$(i) n=1, z_1=1;$$

$$(ii) n > 1, z_n = \lim_{n \rightarrow +\infty} \sum_{i=1}^n (1 + \frac{N}{N - z_{n-1}}) = n;$$

综上

$$z_n = n \quad (11)$$

(2)本算法

设 Z_n 表示采用本算法完成第 n 次分级处理时,扣除重复处理网页后的实际检测网页数。

由式(3)、式(8)和式(9),得检测完成网页数:

$$Z_n = \lim_{n \rightarrow +\infty} (y_{n+1} + \frac{N}{N - y_n}) = n + t_p * x * (n - 1) \quad (12)$$

4.3 效率提升评估

根据上述理论结果,我们对本算法性能提升的大致估计过程如下。

1. 参数 x 和 t_p 估计

(a)单位时间内的网页抓取数 x :根据 JungHoo Cho 等人所发表论文上公布的数据^[14],Stanford WebBase 搜索引擎中每秒可抓取 3000 个网页。这篇文章发表于 2006 年,考虑到搜索技术的发展飞速,我们此时设 $x = 3000\text{page/s}$ 应该是较为保守的。

(b)网页内容分析时长 t_p :在 Rofouei 等人于 2008 年发表了他们在 CPU+GPGPU 并行运算平台上完成人体行为分析算法优化的实验结果^[15]。可以看到,即使在 Intel 3GHz 双核 CPU+带有 128 个 600MHz 运算单元的 NVIDIA GPGPU 运算平台上,对于分辨率仅为 144×192 的视频,平均每幅画面所花时间也达到 4 秒多。考虑到实际网页中往往混合有多张图片、动画甚至视频,网页内容智能分析处理时长应该不会少于 5 秒,因此,我们设 $t_p = 5$ 秒,最终的效率提升评估结果仍然是相对保守的。

2. 算法处理效率比计算

基于上述参数设置和式(12),我们可得到现有第三方分级处理方法和本方法在各种处理时间段内所能完成的网页数,其性能对比如表 1 所列。

表 1 现有第三方分级算法和本算法的网页分级处理效率对比

分级处理次数	现有第三方分级算法处理网页完成数	本算法处理网页完成数	分级处理效率比 (本算法/现有第三方分级算法)
1	1	1	1.000000
10^2	10^2	1485100	14851
10^4	10^4	149995000	14999.5
10^8	10^8	1500099985000	15000.99985

由表 1 知:

(1)即使在对各关键参数进行保守估计的情况下,本算法的处理效率也要远远高于现有第三方分级算法,如本算法 100 次分级处理完成网页数已超过现有第三方分级算法 10000 次的完成数;

(2)仅仅经过 100 次处理,效率提升就已接近 15000 倍,且随着分级处理次数的增加而增长,并无限接近 15001,也即 $(1 + t_p * x)$ 。实际上,该效率提升比正是由算法本身决定的,由式(11)、式(12)两式,可求得提升效率比极限 r :

$$r = \lim_{n \rightarrow +\infty} \frac{Z_n}{n} = \lim_{n \rightarrow +\infty} \frac{(1 + t_p * x) * n - t_p * x}{n} = 1 + t_p * x$$

由上述基于理论分析的万维网内容分级处理效率提升评估可见,通过充分利用由内容关联而形成的万维网社团结构特性,本算法能显著提升现有第三方分级算法的处理效率。

结束语 万维网网络内容因其海量、易变和语义缺失等特性,给万维网内容过滤带来了巨大挑战,现有的第三方分级

算法采用逐个网页处理的方式,效率不高。为此,本文基于 Web 网络社团结构拓扑性质,提出了一种新的 Web 内容分级算法。理论分析证明,相对于现有的第三方分级处理,该算法能显著提升万维网内容分级处理效率;并且本算法框架能同时支持基于人工和基于内容智能分析这两种内容分析方式,具有很好的适应性。

在利用本算法进行万维网内容处理时,某一层次社团结构中那些直接由内容分析处理得到的网页分级标签可能存在不完全匹配的情况,此时可基于该信息对该社团结构做进一步层次划分,保障社团结构和网络内容的相关性,以确保同一万维网社团中网页内容语义标签的一致性和有效性。对此,我们将在后续工作中做进一步研究。

参考文献

- [1] Resnick P, Miller J. PICS: Internet Access Controls Without Censorship[J]. Communications of the ACM, 1996, 39(10): 87-93
- [2] Archer P. ICRAfail-A Lesson For the Future [OL]. 2009. <http://philarcher.org/icra/ICRAfail.pdf>
- [3] Watson G. Internet Content Rating Association (ICRA) Rating System[OL]. <http://256.com/gray/docs/pics/icra.html>
- [4] Hepple M, Ireson N. NLP-enhanced Content Filtering within the POESIA Project[C]// Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC). 2004: 1967-1970
- [5] 国家计算机网络应急技术处理协调中心,中国软件评测中心,等. YDN 138-2006 基于 PC 终端的互联网内容过滤软件技术要求[S]. 北京:人民邮电出版社,2006
- [6] 国家计算机网络应急技术处理协调中心,中国软件评测中心,等. YDN 139-2006 基于 PC 终端的互联网内容过滤软件测试方法[S]. 北京:人民邮电出版社,2006
- [7] Google. We knew the web was big[OL]. 2008. <http://www.oddhubs.com/2012/04/29/we-knew-the-web-was-big/>
- [8] 中国互联网络信息中心. 第 33 次中国互联网络发展状况统计报告[OL]. 2014. <http://www.cnnic.net.cn/hlwzfzyj/hlwxyzbg/hlwtjbg/201403/P020140305346585959798.pdf>
- [9] Yu Li-yang. Introduction to the Semantic Web and Semantic Web Services[M]. New York: Chapman & Hall/CRC, 2007: 6-8
- [10] Flake G W, Lawrence S R, Giles C L, et al. Self-organization and identification of Web communities[J]. IEEE Computer, 2002, 35(3): 66-71
- [11] 邓智龙, 涂文燕. 复杂网络中的社团结构发现方法[J]. 计算机科学, 2012, 39(Z6): 103-108
- [12] Newman M E J, Girvan M. Finding and evaluating community structure in networks[J]. Physical Review E, 2004, 69(2): 1-16
- [13] Shen Hua-wei. Community Structure of Complex Networks [M]. Berlin Heidelberg: Springer-Verlag, 2013: 21-25
- [14] Cho J, Garcia-Molina H, Haveliwala T, et al. Stanford WebBase components and applications[J]. ACM Transactions on Internet Technology (TOIT), 2006, 6(2): 153-186
- [15] Rofouei M, Moazeni M, Sarrafzadeh M. Fast GPU-based space-time correlation for activity recognition in video sequences[C]// IEEE/ACM/IFIP Workshop on Embedded Systems for Real-Time Multimedia (ESTImedia). Los Alamitos, CA, USA: IEEE, 2008: 33-38