计算机研究与发展
Journal of Computer Research and Development
ISSN 1000-1239/CN 11-1777/TP
50(2): 430-436, 2013

基于用户行为的色情网站识别

曹建勋　刘奕群　岑荣伟　马少平　茹立云
(智能技术与系统国家重点实验室(清华大学)　北京　100084)
(清华信息科学与技术国家实验室(筹)　北京　100084)
(清华大学计算机科学与技术系　北京　100084)
(yiqunliu@tsinghua.edu.cn)

Pornography Web Site Identification Based on User Behavior Analysis

Cao Jianxun, Liu Yiqun, Cen Rongwei, Ma Shaoping, and Ru Liyun
(*State Key Laboratory of Intelligent Technology and Systems* (*Tsinghua University*), *Beijing* 100084)
(*Tsinghua National Laboratory for Information Science and Technology*, *Beijing* 100084)
(*Department of Computer Science and Technology*, *Tsinghua University*, *Beijing* 100084)

Abstract　The problem of illegal Web resources, especially pornography sites, poses a major challenge for Web-related applications. Due to the significant differences in page content, site structure and visitors, user behavior patterns on pornography Web sites and ordinary Web sites can be separated from each other. With the help of a popular commercial search engine in China, large scale user behavior data is collected and it is found that when users surf in porn sites, their behaviors are significantly different from that when they are visiting ordinary Web sites. These differences in user behavior patterns can help us separate porn sites from other ones. A number of behavior features are proposed and combined with machine learning algorithms to develop a porn site identification method. Experimental results show effectiveness of the proposed method.

Key words　pornography site; illegal Web resources; user behavior analysis; search engine; Web browsing

摘　要　以色情网站为代表的万维网非法资源已经成为互联网应用普及过程中的重大挑战.由于色情网站与普通网站的内容特征、结构形式和访问者群体都有显著的差异,这造成了用户对色情网站和普通网站的访问行为的差异.在某商业搜索引擎的协助下,收集了海量规模互联网用户访问日志,基于对日志中所记载用户行为的挖掘,验证了用户访问色情网站与普通网站时的行为确实具有明显的差异.基于此类差异设计了一系列用户行为特征,并结合机器学习方法,设计了基于用户行为的色情网站识别方法.实验表明,该方法可以较准确、高效地从网站中识别色情网站.

关键词　色情网站;网络非法资源;用户行为分析;搜索引擎;网络浏览

中图法分类号　TP391

收稿日期:2010-11-25;**修回日期**:2011-12-16

基金项目:**国家"八六三"高技术研究发展计划基金项目**(2011AA01A205);**国家自然科学基金项目**(60903107,61073071);**高等学校博士学科点专项科研基金项目**(20090002120005)

通信作者:**刘奕群**(yiqunliu@tsinghua.edu.cn)

我国正处于互联网高速发展的浪潮之中,一方面,截止 2011 年底,中国网民规模达到 4.85 亿,位居世界首位,网页数量达到 600 亿以上,自 2003 年以来已经保持了多年的快速增长;另一方面,我国人口中的互联网普及率目前为 36.2%,仅略高于世界平均水平,远低于西方发达国家的水平. 互联网普及率相对较低的现状与"信息下乡"、手机上网普及等推动因素的存在,将使我国互联网发展水平在可预见的将来继续保持高速增长[1].

然而,在互联网高速发展的今天,网络的负面效应和负面信息也逐渐膨胀起来,其中以网络欺诈、网络色情、网络违法信息等为代表的以提供非法内容为主的网站赢利最高,对网民群体造成的影响也最为严重. 根据《Internet Filter Review》的统计结果表明,目前全球大约有 420 万个色情网站,占所有网站总数的 12%,互联网上色情网页超过了 3.72 亿个[2].

当前,色情网站的识别工作已取得了一定的进展,如 Lee 等人提出的多个内容分级体系[3]、苏贵洋等人提出的用于中文色情文本过滤的 KNN 构造算法[4]等. 这些工作对色情网站的识别多数基于页面内容和网站的链接结构,缺乏一定的灵活性,对特殊结构的色情网站也会失去识别的准确性.

相比已有的色情网站识别工作,我们提出一种基于用户行为分析的色情网站识别方法. 一方面,网络用户的行为信息对于网络应用系统研究而言是最可宝贵的反馈信息,用户行为分析也是在网络应用研究(如搜索引擎系统设计、网络服务系统设计)中发挥重要作用的各种算法(如 PageRank 算法)的基本依据之一;另一方面,由于非法网页的创建目的、页面内容和组织形式都与普通页面有一定的差别,用户访问色情网站的行为模式也必然与其访问普通网站时有较大的差异.

1 相关工作

由于西方法律法规监管方式的不同,相当比例的网络色情内容识别工作是以建立内容分级制度为目的,因而由网站自身或第三方机构对网站内容进行评级就为色情内容识别工作提供了很大的便利,如 Lee 等人提出的多个内容分级系统就都将这种评级信息作为重要的分类依据. 除此之外,多媒体领域的研究人员(如 Arentz 等人和 Zheng 等人)[5-6]提出了多种基于形状或颜色特征的色情多媒体资源内容识别方法;Hammami 等人[7]也尝试融合包括内容特征、网页结构特征和多媒体内容特征在内的多种特征对色情资源加以识别. 在国内,色情网站的识别工作大多数则是依靠对色情文本的识别而达到最终识别网站的方法. 如苏贵洋等人提出的用于中文色情文本过滤的 KNN 构造算法等等.

这部分工作能够对网络中存在的某些非法资源起到较好的识别效果,但面对网络中存在的大量非法资源网页,已有识别算法在设计思路上存在着两个难以逾越的技术难题:

1)识别方法的效率问题. 当前大多数针对色情内容的识别工作都需要涉及对于多媒体内容的分析与处理,这种识别方法在进行海量规模网络数据的处理时将会面临巨大的挑战. 而针对欺诈内容的识别往往也需进行大量网站内容与结构特征的相互比对,这同样会造成识别算法效率低下的问题.

2)识别方法的通用性问题. 当前针对特定非法内容设计专用识别方法的识别思路,造成这类识别方法往往只能处理某种特定的非法内容. 一旦非法内容更换网页结构或内容关键词,就会造成现有识别方法失效的情况. 而针对海量非法内容实时更新的现实,当前的识别方法往往应接不暇.

2 色情网站的用户行为特征

如何基于用户行为来识别色情网站是我们关注的重点. 随着搜索引擎技术的发展,由搜索引擎公司提供的浏览器工具栏越来越为广大网络用户所接受. 浏览器工具栏可以为用户提供直接的搜索引擎访问接口,同时也可以提供弹出窗口过滤、下载加速、网络书签等多种附加功能. 主流搜索引擎公司如谷歌(http://toolbar.google.com/)、雅虎(http://toolbar.yahoo.com/)、百度(http://bar.baidu.com/)、微软(http://toolbar.live.com/)等都推出了自己的浏览器工具栏服务,不少公司还把工具条与其他软件产品捆绑发行以加强推广. 与此同时,大多数搜索引擎供应商也通过工具栏基于匿名策略收集用户的 Web 访问行为数据,以便为工具栏用户提供更多个性化的增值服务. 最近,一些研究人员也开始利用这部分 Web 访问行为数据对网络用户的行为特征加以研究和利用. 如今,基于互联网访问日志的用户行为分析被广泛应用于搜索引擎算法改进[8-10]、竞价广告投放[11-12]、作弊页面识别[13]等方面的研究中. 在某商业搜索引擎的帮助下,我们获取了海量规模的网络访问日志(数据详情将在第 4 节叙述),通过这部分日志进行数据挖掘,我们可以发现普通网站

和色情网站不同的一些行为模式,提取能够描述这种行为模式差异的用户行为特征,并通过机器学习的方式来试图将两类网站区别开来.在本文中,我们提出了下面的4个特征.

2.1 特定时间段网站访问率

对于一个网站来说,访问量是作为网站知名度的一个重要评价标准.在互联网上,用户的每一次网站访问都通过日志的形式保存下来,同时被记录的还有用户的访问时间.考虑色情网站,由于很多用户会选择在夜间访问;同时,很多用户同样不会选择在上网高峰期浏览色情网站,因此色情网站在网络访问时间上会与大多数网站有比较明显的区别.

这样,我们考虑通过统计不同时间段网站访问的频率用来作为辨别色情页面的一个重要特征.因此定义时间段$[s,e]$间网站访问率$VP_{s,e}(p)$如下:

$$VP_{s,e}(p)=\frac{s\text{到}e\text{时间内网站访问量}}{\text{网站总访问量}}. \quad (1)$$

如果特别考察和$VP_{0,7}(p)$这两个参数,会得到图1和图2所示的统计结果.很容易发现,色情网站和非色情网站在数据统计中呈现出不同的分布趋势.

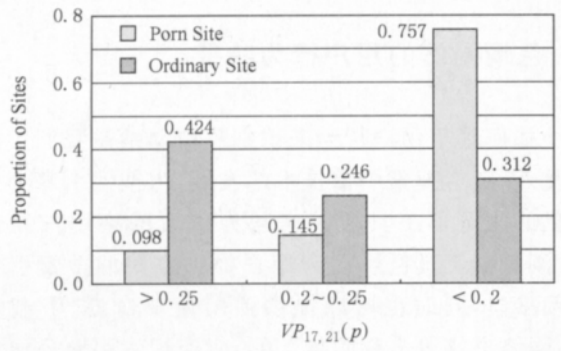

Fig. 1 $VP_{17,21}(p)$ distribution of porn sites and non-porn sites.

图1 色情网站和非色情网站的$VP_{17,21}(p)$分布

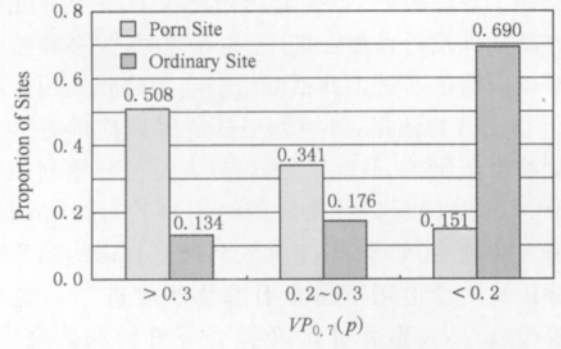

Fig. 2 $VP_{0,7}(p)$ distribution of porn sites and non-porn sites.

图2 色情网站和非色情网站的$VP_{0,7}(p)$分布

我们可以看出,色情网站$VP_{0,7}(p)$的平均值要大于相应非色情网站的平均值,而色情网站的$VP_{17,21}(p)$的平均值要小于相应非色情网站的平均值.

2.2 搜索引擎及网站内部网页访问率

用户访问一个网站可能通过这样几种方式:搜索引擎查询结果访问、朋友或可信广告推荐、书签或历史访问、感兴趣的超链接点击等.对于色情网站,通过搜索引擎查询结果的访问最为常见;同时,由于用户访问色情网站时普遍停留时间比较长,色情网站内部访问比率也相对比较高.

因此,我们定义搜索引擎及网站内部页面访问率$SEIOV(p)$为

$$SEIOV(p)=\frac{\text{搜索引擎访问量}+\text{内部网页访问量}}{\text{网页总访问量}}. \quad (2)$$

事实上,如果在式(2)分子中增加色情网页之间的访问量,则色情网站的$SEIOV(p)$值将非常接近1,这在调研的初步统计中也有所展示.但由于考虑色情网页之间的访问量时容易造成将计算$SEIOV(p)$复杂化,因此不将其计算.通过实验,我们得到图3所示的数据:

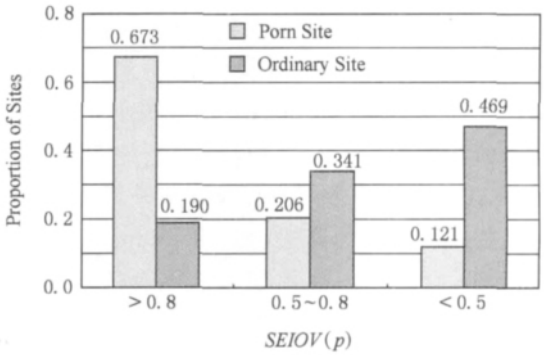

Fig. 3 $SEIOV(p)$ distribution of porn sites and non-porn sites.

图3 色情网站和非色情网站的$SEIOV(p)$分布

从数据可以看出,色情网站的值普遍高于相应非色情网站的值,这很有助于我们去通过特征分辨色情网站.

2.3 色情词关联度

几乎每一个网站的访问日志都会包含通过搜索引擎访问的部分.虽然大多数商业网站都会尽量去推广自己的网站名称来让自己的网站知名度更高,从而带来更大的效益,但对于色情网站,政府的监管制度使得这些网站必须通过非正当宣传手段使得用户去访问.这时,色情网站便利用搜索引擎为用户提供一个很便利的入口.

因此,一个网站的访问是否包含用户通过搜索引擎查询色情词而产生的点击行为可以比较好地判定一个网站有可能为色情网站.一个容易接受的事实就是如果用户通过查询色情词而引导至一个网站,那么这个网站极有可能是色情网站.

利用这样一个事实,我们定义色情词关联度$SWC(P)$.现在需要考虑这个参数的取值问题.由于很多色情词都或多或少包含歧义,这使得不能简单地将该参数二值化.由于色情词本身就有分级,为简单起见,将查询词分为3级:纯色情词、非色情词、歧义色情词.因此,不妨对网站考虑如下的分级策略:

1) 对于包含纯色情词关联的网站标记为1;

2) 对于无任何纯色情词或歧义色情词关联的网站标记为0;

3) 对于包含歧义色情词关联却不包含纯色情词关联的网站,计算色情词关联度函数如下:

$$SWC(p)=\frac{\text{网站的歧义色情词关联次数}}{\text{网站通过搜索引擎访问次数}}. \quad (3)$$

容易看出,$SWC(P)$是一个0~1之间的实数,且不会由于网站的访问量大而导致该参数的价值失效.通过实验我们得到了图4的数据.从图4我们可以看出色情网站与非色情网站在该特征上差别非常显著:

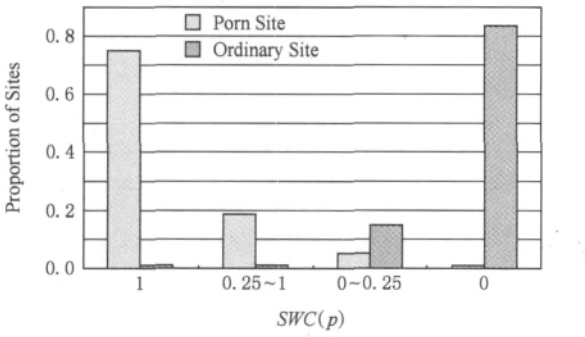

Fig. 4 $SWC(p)$ distribution of porn sites and non-porn sites.

图4 色情网站和非色情网站的$SWC(p)$分布

在计算色情词关联度时,我们需要用到一个已分级的色情词表.虽然我们在某商业搜索引擎的协助下获取到一个色情词表,但其中没有色情分级信息,而且由于需要动态地计算某些新出现的网站的色情词关联度,因此色情词表也需要动态更新,用来防止色情词表中太多的冗余项和遗漏项.因此,我们需要一种采用基于用户行为方法提取的色情词.

实验表明,色情词的搜索行为特征相比于普通查询词有很大差别.在商业搜索引擎的协助下,我们获取了海量规模的搜索引擎用户行为日志(数据详情将在第4节叙述).经过分析,我们发现色情词的特殊性主要表现在以下几个方面:

1) 多数色情词查询词翻页量比较多.

用户查询色情词时很难会找到一个查询项会立刻终止查询.用户在查询色情词时将非常有可能点击到查询结果的很多页.这些行为都记录在搜索引擎日志中,因此可以挖掘查询词的最大翻页数作为识别色情查询词的一个特征.

2) 多数色情查询词的查询会话(session)比较长.

一个查询词的会话定义为:在某个特定用户查询特定查询词时,用户在该查询词中最后一次动作行为与第1次动作行为之间的时间长度记为一个会话.用户在浏览查询词结果时,由于会在色情网站停留比较久的时间,而在浏览完网站后很可能继续点击下一个搜索项,因此大多数色情查询词会话会比普通查询词的会话长.

3) 色情查询词对应的用户会话中用户搜索交互行为较多.

在搜索引擎访问日志中,用户在查询某个查询词时的各种搜索交互行为均会被记录.我们发现除查询与结果点击行为之外,用户在查询色情词时与搜索引擎其他方面的交互内容也较多,如切换至视频搜索、点击广告、点击站内信息、点击网页快照、点击查询提示、点击帮助或意见反馈等.

这些特征都会间接地反映了一个查询词的性质.经过实验表明,这些特征会在一定程度上区分色情查询词与普通查询词.

在面向海量搜索引擎访问日志提取了上述特征之后,我们考虑使用聚类方法提取色情查询词.由于当特征比较显著且足够时,不同聚类方法得到的聚类结果是基本相同的.考虑到效率问题,我们采用了速度较快的K-Means算法.

在实验中我们发现,虽然选取了不同的聚类数K和不同的随机种子,但在得到的聚类结果中有一个数量几乎恒定地包含了较多的色情词的类,我们称之为色情词母类.为了提高聚类的准确率,我们对色情词母类采用同样的K值和种子集合进行了二次聚类,这样得到了一个相对比较好的色情词集合.通过进行抽样计算,该色情词类的召回率达到了81.6%,通过对该色情词集合进行分级标注,即可用于计算色情词关联度的参数.

2.4 URL 特定字符串关联度

网站名对于用户访问有很强的指导作用. 而大多数色情页面也通过网站名对用户起到一定的暗示作用. 通过检验网站名中是否包含一些特定的字符串序列(如 bo mm av bb cao se pp xx sao jj 等),可以有效地辨别部分色情页面. 经过统计,当网站名包含如表 1 中所列的字符串序列时,该网站很大程度上可能是色情页面:

Table 1 Some String Sequence May Be Used to Distinguish the Name of Porn Sites

表 1 某些可能用于区别色情网站名称的字符串序列

String Sequence	String Sequence
bo	se
mm	pp
av	xx
bb	sao
cao	jj

3 色情网站识别流程

在提取了网站的相关特征之后,我们将用机器学习算法对网站进行分类. 由于特征数目相对比较少,同时特征也比较明显,因此将采用朴素贝叶斯和决策树两种方法尝试对色情网站进行识别. 我们设计了图 5 的流程进行识别实验:

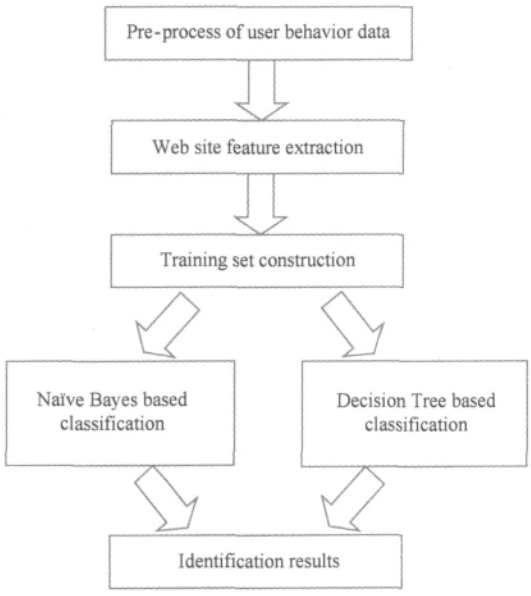

Fig. 5 Flowchart of identifying porn sites.

图 5 色情网站识别流程图

4 实验结果与讨论

我们采用了从 2009-11-01 至 2009-11-30 这一个月的数据作为用来分析的数据源. 该数据源包括两部分:一是海量的网络访问日志,这部分日志记录了用户的访问时间、用户 IP、浏览器产生的 ID、目的 URL 和源 URL 这 5 部分;二是海量的搜索引擎日志,这里我们采用搜狗搜索引擎提供的这一个月的搜索日志. 搜索日志记录了大量的用户在搜索引擎页面产生的动作行为,包括点击行为、鼠标滚轮行为等.

通过实验我们发现,这一个月的数据中大约有 1 262.9 万的网站访问记录,但月访问量超过 300 的网站只有大约 11.6 万个. 我们在后面抽样时将从这些月访问量超过 300 的网站中选取.

我们从月访问量超过 300 的网站中抽取大约 10%的网站进行标记,取其中全部的可访问的非色情网站 3 331 个;同时,由于标注出的可访问的色情网站数量相对比较小,因此加入了一部分已知的色情网站,这样共有色情网站 2 399 个.

Table 2 The Proportion of Porn Sites and Non-Porn Sites in Sample

表 2 样本中色情网站与非色情网站的比例

Web Site Type	Site Number	Percentage/%
Pornography Site	2 399	41.87
Ordinary Site	3 331	58.13
Total	5 730	100

在测试集方面,我们并没有使用单独的测试集,而为了保证生成模型的准确性而不出现过拟合现象,我们来选择 10 折交叉验证来选择和估计模型性能.

首先采用贝叶斯分类方法对测试网站进行分类,得到如表 3 所示的分类结果:

Table 3 Naïve Bayes Classifier Classification Results

表 3 朴素贝叶斯分类器分类结果

Classified Instances	Amount	Percentage/%
Correctly Classified	5 592	97.59
Wrongly Classified	138	2.41
Total	5 730	100

然后采用决策树分类方法对测试网站进行分类,得到如表 4 所示的分类结果:

Table 4 Decision Tree Classifier Classification Results

表 4 决策树分类器分类结果

Classified Instances	Amount	Percentage/%
Correctly Classified	5 598	97.70
Wrongly Classified	132	2.30
Total	5 730	100

我们进一步计算了两种分类方法识别色情网站的准确率和召回率,如表 5 所示:

Table5 The Precision and Recall of Two Kinds of Classification Method to Identify Porn Sites

表 5 两种分类方法识别色情网站的准确率和召回率 %

Evaluation Metrics	Naïve Bayes	Decision Tree
Precision	96.66	96.15
Recall	97.62	97.83
F-measure	97.13	96.98

从朴素贝叶斯和决策树两种方法的分类结果来看,准确率、召回率和 F-measure 均较高且相差不大,一方面说明了方法的正确性和可行性,另一方面也说明了在特征比较显著地情况下,使用哪一种分类方法得到的效果基本相同.

单独考察分类错误的实例,我们发现这部分网站主要是一些访问量很低、页面内部点击量比较小的网站,且大多数的页面已经无法访问,使得大部分特征与同类网站的特征有所差别,导致分类结果错误.但由于错误的实例非常少,且均非大中型网站,因此这部分错误对结果的影响可以忽略不计.

5 结论及未来工作

大多数的色情网站识别算法都是基于页面内容和网站的链接结构.我们提出了一种新的基于用户互联网行为的色情网站提取方法.通过分析海量规模的网络访问日志,发现了一些色情网站与普通网站之间不同的行为模式,提取了一些用于识别色情网站的特征参数.基于这些特征参数,我们结合机器学习的方法并对网站进行了分类识别.通过这种方法我们有效地从普通网站中分辨出色情网站,并在准确率上有了较大的提高.

在今后的工作中,我们希望能将这种方法推广,即采用类似的思想和方法识别网络中更多类型的非法资源,如赌博网站、欺诈网站等.

参 考 文 献

[1] China Internet Network Information Center. The 28th statistical report on Internet development in China [R]. Beijing: China Internet Network Information Center, 2011 (in Chinese)
(中国互联网络信息中心(CNNIC). 第 28 次中国互联网络发展状况统计报告[R]. 北京: 中国互联网络信息中心, 2011)

[2] TechMediaNetwork. Internet filter software reviews 2011 [R]. Ogden, Utah: TechMediaNetwork. 2011

[3] Lee L, Luh C. Generation of pornographic blacklist and its incremental update using an inverse chi-square based method [J]. Information Processing and Management, 2008, 44(5): 1698-1706

[4] Su Guiyang, Li Jianhua, Ma Yinghong, et al. A KNN algorithm on Chinese erotic text filtering [J]. Journal of Shanghai Jiaotong University, 2004, 38: 76-79 (in Chinese)
(苏贵洋, 李建华, 马颖红, 等. 用于中文色情文本过滤的近邻法构造算法[J]. 上海交通大学学报, 2004, 38: 76-79)

[5] Arentz W A, Olstad B. Classifying offensive sites based on image content [J]. Computer Vision and Image Understanding: Special Issue on Color for Image Indexing and Retrieval, 2004, 94(1-3): 295-310

[6] Zheng Q F, Zeng W, Wen G, et al. Shape-based adult image detection [C] //Proc of the 3rd IEEE Int Conf on Image and Graphics. Piscataway, NJ: IEEE, 2004: 150-153

[7] Hammami M, Chahir Y, Chen L. WebGuard: A Web filtering engine combining textual, structural, and visual content-based analysis [J]. IEEE Trans on Knowledge and Data Engineering, 2006, 18(2): 272-284

[8] Liu Y, Gao B, Liu T, et al. BrowseRank: letting Web users vote for page importance [C] //Proc of the 31st ACM SIGIR Conf. New York: ACM, 2008: 451-458

[9] Bilenko M, White R W. Mining the search trails of surfing crowds: Identifying relevant websites from user activity [C] //Proc of the 17th Int Conf on World Wide Web. New York: ACM, 2008: 51-60

[10] Liu Y, Zhang M, Ma S, et al. User browsing graph: Structure, evolution and application [C] //Proc of the 2nd ACM Int Conf on Web Search and Data Mining. New York: ACM, 2009: 1-4

[11] Chen Lei, Liu Yiqun, Ru Liyun, et al. Performance evaluation of online sponsored search based on user log analysis [J]. Journal of Chinese Information Processing, 2009, 22(6): 92-97 (in Chinese)
(陈磊, 刘奕群, 茹立云, 等. 基于用户日志挖掘的搜索引擎广告效果分析[J]. 中文信息学报, 2009, 22(6): 92-97)

[12] Wang Jiazhuo, Liu Yiqun, Ma Shaoping, et al. Sponsored search performance analysis based on user behavior information [J]. Journal of Computer Research and Development, 2011, 48(1): 133-138 (in Chinese)

(王家卓, 刘奕群, 马少平, 等. 基于用户行为的竞价广告效果分析[J]. 计算机研究与发展, 2011, 48(1): 133-138)

[13] Liu Yiqun, Cen Rongwei, Zhang Min, et al. Identifying Web spam with user behavior analysis [C] //Proc of the 4th Int Workshop on Adversarial Information Retrieval on the Web. New York: ACM, 2008: 9-16

Cao Jianxun, born in 1990. Received his bachelor's degree from Tsinghua University in 2010, and now PhD candidate in Tsinghua University. His current research interests include network engineering and user behavior analysis(hailercao@qq.com).

Liu Yiqun, born in 1981. Received his bachelor and PhD degrees from Tsinghua University in 2003 and 2007, respectively. He is now assistant professor in the Department of Computer Science and Technology in Tsinghua University. Member of China Computer Federation. His current research interests include Web information retrieval, Web user behavior analysis and machine learning.

Cen Rongwei, born in 1984. Received his bachelor and PhD degrees in Tsinghua University in 2005 and 2010, respectively. His current research interests include Web search engine and search performance evaluation(cenrongwei@gmail.com).

Ma Shaoping, born in 1961. Received his bachelor and PhD degrees in Tsinghua University. He is now professor and PhD supervisor in the Department of Computer Science and Technology in Tsinghua University. Member of China Computer Federation. His current research interests include Web information retrieval, Web user behavior analysis and machine learning(msp@tsinghua.edu.cn).

Ru Liyun, born in 1979. Received his bachelor's degree from Tsinghua University in 2002 and now PhD candidate in Tsinghua University. His current research interests include natural language processing and user behavior analysis(ruliyun@sogou-inc.com).