

# A complementary note on "Web log analysis method"

By Hamid R. Jamali

Published in: *Informology*, 2004, vol.1 No. 3, pp. 177-185.

**Abstract:** This short article is a complementary note on the article "Web log analysis method" by Hajar Sotudeh published in *Informology* (vol.1, no.2, pp.69-84) which discussed methodological aspects of the Web log analysis method. The present note explains some of the technical problems of the method including caching, user identification, session continuity, effects of spiders and robots, size of data and disadvantages of proprietary web log analysis softwares. Some techniques for improving the quality of research output of the web log analysis are also described.

## یادداشتی بر روش تحلیل گزارش های وب (وب لاگ) نویسنده: حمید رضا جمالی مهمونی اطلاع شناسی. جلد اول، شماره سوم. صص. ۱۷۷-۱۸۵.

### چکیده

این نوشته یک یادداشت تکمیلی است بر مقاله ای تحت عنوان "روش تحلیل گزارش های وب (وب لاگ)" به قلم هاجرستوده که در صفحات ۶۹-۸۴ شماره دوم از جلد اول نشریه *اطلاع شناسی* منتشر شده و ضمن معرفی روش تحلیل گزارش های وب به تشریح ویژگی های روش شناختی آن پرداخته است. یادداشت حاضر با تمرکز بر مسائل فنی به شرح برخی مشکلات موجود در این روش از جمله تاثیر حافظه پنهان و روبات ها، مشکل حجم عظیم داده ها، پیچیدگی شناسایی کاربران و طول نشست و نقص نرم افزارهای تجاری موجود برای تحلیل گزارش ها می پردازد. همچنین نوشته به برخی شیوه های ممکن برای بهبود کیفیت این روش تحقیق برای سنجش رفتار اطلاعاتی کاربران اشاره می کند.

### مقدمه

در صفحات ۶۹-۸۴ شماره دوم از جلد اول نشریه *اطلاع شناسی* مقاله ای تحت عنوان "روش تحلیل گزارش های وب (وب لاگ)"<sup>۱</sup> به قلم هاجرستوده منتشر شد که اولین مقاله در این موضوع در متون فارسی حوزه کتابداری و اطلاع رسانی محسوب می شود. اگرچه ممکن است متخصصین حوزه علوم رایانه ای در ایران تا به حال پژوهش هایی را بر روی گزارش های وب انجام داده باشند که اغلب با رویکرد فنی خاص این متخصصین و جهت سنجش کارایی وب سایت ها انجام می شود، اما تا به حال پژوهشی با این روش با رویکرد علم اطلاع رسانی در ایران انجام نشده است. صرف نظر از دلایل فنی که این عدم پژوهش می تواند داشته باشد از جمله مشکل دسترسی به فایل های لاگ که در اختیار صاحبان خدمت دهنده های اینترنت قرار دارند و یا نبود همکاری علمی لازم میان متخصصین رایانه و کتابداران، یقیناً یکی از دلایل این نقیصه نبود دانش و آشنایی کافی با این روش در میان کتابداران و اطلاع رسانی ایرانی است. از این رو مقاله مذکور می تواند گام بسیار مثبتی در جهت معرفی این روش باشد. از آنجا که هدف مولف بیشتر بررسی نقاط قوت و ضعف این روش به لحاظ روش شناسی {پایان صفحه ۱۷۷} بوده است (ص ۷۰)، برخی از نکات به ویژه مسائل فنی تنهابه صوت گذرا اشاره شده و برخی نواقص از قلم افتاده است. نوشته حاضر سعی دارد تا جهت تکمیل مقاله ارزشمند مذکور نکاتی را به ویژه با تکیه بر تجربیات گروه تحقیقاتی سایبر<sup>۲</sup> که از گروه های تحقیقاتی فعال در به کارگیری این روش است، به صورت تکمیلی مطرح کند.

<sup>1</sup> Web Log

<sup>2</sup> Centre for Information Behaviour and the Evaluation of Research: Ciber (City University London)

فایل های گزارش وب یا به زعم محققین گروه سایبر "تلویزیون های مدار بسته فضای سایبر"<sup>۳</sup> (نیکلاس<sup>۴</sup> و دیگران، ۱۹۹۹، ص ۲۶۳)، فایل های متنی هستند که در حالت پیش فرض در شاخه اصلی یک سایت یعنی همان شاخه ای که صفحه خانگی یک سایت ذخیره شده ذخیره می شوند (گوتزمن<sup>۵</sup>، ۱۹۹۹). این فایل ها در حقیقت در ابتدا به عنوان یک محصول فرعی پروتکل انتقال فرامتن<sup>۶</sup> ایجاد شدند (فیبیر<sup>۷</sup>). استفاده از این فایل ها برای بررسی رفتار اطلاع یابی کاربران و کارایی وب سایت ها که به نوعی تداوم تحقیقات مشابه دهه های قبل روی گزارش های تراکنشی<sup>۸</sup> فهرست های رایانه ای و نظام های ذخیره و بازیابی محسوب می شود به لحاظ مشکلات و ابهاماتی که به همراه دارد با دیدگاه های ضد و نقیضی همراه بوده است. برخی چون گوتزمن آنها را به "گنجینه اطلاعات ارزشمند" تشبیه کرده، برخی چون فولر و دو گراف<sup>۹</sup> با تردید آنها را منشا نتایج گمراه کننده در مورد فعالیت کاربران دانسته و برخی دیگر نظیر اودل<sup>۱۰</sup> از این هم فراتر رفته و آنها را "دروغهای لعنت شده" نامیده اند (نیکلاس و هانتینگتون<sup>۱۱</sup>، ۲۰۰۳، ص ۳۹۳).

### مشکلات روش تحلیل گزارش های وب

نویسنده مقاله در صفحه ۷۳ به اجمال به برخی مشکلات روش از جمله وجود حافظه پنهان<sup>۱۲</sup>، عدم شناسایی کاربر، حجم زیاد داده ها و عدم ثبت برخی وقایع نظیر استفاده از کلیدهای Forward و Back در فایل های گزارش اشاره کرده است. در اینجا ضمن تشریح بیشتر این موارد به برخی موارد نیز دیگر اشاره میشود.

**الف. حافظه پنهان.** وجود حافظه پنهان تاثیر چشمگیری بر تحلیل آماری نتایج حاصل از تحلیل گزارشها دارد و موجب افت آمار مربوط به استفاده از صفحات وب می شود. این مشکل چند منشا دارد. بخشی از آن ناشی از عملکرد نرم افزار مرورگر مورد استفاده توسط کاربر است. این نرم افزارها به عنوان مثال اینترنت اکسپلورر<sup>۱۳</sup> جهت افزایش کارایی خود و تسریع فرایند مرور صفحات توسط کاربر، اقدام به ذخیره صفحات رویت شده طی یک دوره زمانی خاص بر روی فضای محدودی از رایانه کاربر {پایان صفحه ۱۷۸} می کنند. در طول این دوره اگر کاربر صفحه ای را بارگذاری کند، برای بار دوم نیازمند به بارگذاری آن از روی اینترنت نیست چرا که مرورگر در دفعات بعد آن را از آرشیو موقت خود فرامی خواند نه از روی اینترنت. عدم ثبت استفاده از کلیدهای Forward و Back نیز به دلیل همین ویژگی است. و به همین دلیل است که گاه برای اطمینان از روزآمد بودن صفحات وب، کاربر نیاز پیدا می کند تا از کلید بارگذاری مجدد یا Refresh استفاده کند. البته کاربر می تواند این ویژگی مرورگر را غیر فعال و یا به دلخواه خود تنظیم کند. بنا بر تحقیق فیبر ۳۵ تا ۵۵ درصد از عملکرد کاربر به دلیل همین ویژگی مرورگرها در فایل های گزارش ثبت نمی شود. ذکر این نکته نیز لازم است که بنا بر تحقیق نیکلاس و همکارانش (۲۰۰۲، ص ۶۳) میزان ذخیره صفحات در حافظه پنهان رایانه کاربر تا حدودی متاثر از معماری و ساختار وب سایت است.

نوع دیگر حافظه پنهان، خدمت دهنده های پروکسی<sup>۱۴</sup> هستند. پروکسی ها همانند واسطه ای میان کاربر و خدمت دهنده در جهت افزایش کارایی و فیلتر کردن اطلاعات عمل میکنند. پروکسی ها گاه جهت افزایش سرعت شبکه کاربران خود اقدام به ذخیره صفحات وب می کنند. در نتیجه زمانی که کاربر صفحه ای را درخواست می کند، صفحه از حافظه پروکسی به کاربر منتقل می شود نه از خدمت دهنده اصلی وب و طبیعتاً این درخواست کاربر در گزارش وب خدمت دهنده اصلی ثبت نمی شود.

وجود حافظه پنهان نه تنها بر ثبت آمار استفاده از سایت اثر دارد بلکه در تحلیل زمان رویت هر صفحه توسط کاربر نیز تاثیر منفی دارد (نیکلاس و دیگران، ۲۰۰۲، ص ۶۳). چرا که کاربر ممکن است چندین صفحه را بدون اینکه در فایل گزارش ثبت شود رویت کند و در نتیجه محقق در زمان تحلیل گزارش چنین تصور کند که کاربر تمام این مدت را صرف مشاهده یک صفحه یعنی آخرین صفحه رویت شده توسط وی که در گزارش ثبت شده، کرده است.

<sup>3</sup> CCTV of Cyberspace

<sup>4</sup> Nicholas

<sup>5</sup> Gutzman

<sup>6</sup> HyperText Transfer Protocol: HTTP

<sup>7</sup> Fieber

<sup>8</sup> Transactional Log

<sup>9</sup> Fuller & De Graaf

<sup>10</sup> Udell

<sup>11</sup> Huntington

<sup>12</sup> Cache

<sup>13</sup> Internet Explorer

<sup>14</sup> Proxy Server

ب. روبات<sup>۱۵</sup> ها و خزنده های اینترنت<sup>۱۶</sup>. یکی دیگر از عوامل گمراه کننده در این روش، وجود خزنده ها و روبات هایی است که به صورت مرتب به کاوش وب میپردازند. نمونه بارز آنها روبات هایی هستند که توسط موتورهای جستجو جهت بررسی محتوای صفحات وب و نمایه سازی آنها به خدمت گرفته می شوند. هر بار که چنین روباتی یک سایت را کاوش میکند تمامی عملکرد آن همانند عملکرد یک کاربر ثبت می شود در حالی که در حقیقت هیچ انسانی از سایت استفاده نکرده است. به زعم {پایان صفحه ۱۷۹} گوترمن (۱۹۹۹) حتی تا یک سوم از ترافیک ثبت شده یک وب سایت ممکن است ناشی از عملکرد روبات ها باشد. البته در مواردی کاوش سایت توسط روبات ها به دلیل اینکه در فایلی به نام "Robot.txt" روی خدمت دهنده وب ثبت می شود، قابل شناسایی است و می توان با رجوع به آنها این نوع اطلاعات را از گزارشهای وب حذف کرد. در عین حال تعداد رو به افزایش روبات های ناشناس وب نیز مشکل دیگری است (نیکلاس و دیگران، ۲۰۰۲، ص ۶۴)

ج. حجم زیاد داده ها. از آنجا که تمام وقایع با جزئیات در فایل گزارش ثبت می شوند، حجم داده های موجود در این فایل ها بسیار زیاد است. به عنوان نمونه استفاده از سایت روزنامه های The Times و The Sunday Times روزانه منجر به ثبت حدود ۱۷۴ میلیون سطر داده در فایل گزارش آنها می شود. (نیکلاس و دیگران، ۲۰۰۰ ص ۴۰۲). طبیعتاً تحلیل صحیح این حجم از داده ممکن است با مشکلاتی مواجه شود. اما در حقیقت تمامی این داده ها مفید نیستند. بسته به نوع وب سایت و محتوای آن بخش قابل توجهی از این داده ها می تواند ناشی از ثبت انتقال اجزاء یک صفحه (نظیر تصاویر و آیکون ها) به کاربر باشد که ارزش اطلاعاتی ندارند. در تحقیقی بر روی دو سایت مربوط به اطلاعات بهداشتی حدود ۸۵ درصد از حجم فایل های گزارش، مربوط به این نوع داده های زاید تشخیص داده شده و حذف شد (نیکلاس و دیگران، ۲۰۰۲، ص ۶۶). بسته به اهداف تحقیق ممکن است لازم باشد محقق بخش قابل توجه دیگری از داده ها را نیز از تحقیق حذف کند. به عنوان مثال در تحلیل گزارش وب یک نشریه الکترونیکی که با فرمت PDF عرضه می شود، می توان تنها به سطوری از داده ها پرداخت که مربوط به این نوع فایل ها هستند.

د. شناسایی کاربر. مطالعه رفتار کاربر بدون آنکه شناخت دقیقی از هویت، منشا و ویژگی های کاربر داشته باشیم حاصل چندان نخواهد داشت. یکی از بزرگترین چالش ها در روش تحلیل گزارشهای وب مشکل شناسایی کاربر است. این مشکل دلایل متعدد دارد. گزارشهای وب از پروتکل اینترنت<sup>۱۷</sup> یا آی پی برای شناسایی کاربر استفاده می کنند. هر رایانه ای بر روی اینترنت دارای یک آدرس منحصر به فرد است که آدرس آی پی نام دارد. این آدرس (به عنوان مثال ۱۳۲،۱۵۱،۳،۹۰) از چهار عدد تشکیل شده که هر یک در محدوده صفر تا ۲۵۵ قرار دارند (آرمس<sup>۱۸</sup>، ۲۰۰۰، ص ۲۳). زمانی که کاربری از وب استفاده میکند آدرس آی پی رایانه وی در گزارش ثبت می شود. {پایان صفحه ۱۸۰} اولین مشکل از همین جا ناشی می شود چرا که در حقیقت فایل گزارش عملکرد یک رایانه را ثبت می کند نه یک کاربر مشخص واحد را. چه بسا چندین نفر از یک رایانه به صورت مشترک استفاده کنند نظیر رایانه های عمومی داخل کتابخانه ها یا خدمات عمومی یا کافه های اینترنتی یا حتی رایانه مشترک اعضا یک خانواده. مشکل دوم این است که دو نوع آدرس آی پی وجود دارد: ثابت و پویا یا شناور. به دلیل محدود بودن آدرس های آی پی، بسیاری از فراهم کنندگان خدمات اینترنتی<sup>۱۹</sup> یا آی.اس.پی ها از آدرس های آی پی شناور برای مشترکین خود استفاده میکنند. به این صورت که هر بار که مشترک اقدام به اتصال به اینترنت از طریق آی.اس.پی می کند، آی.اس.پی یک آدرس شناور به وی اختصاص می دهد که به احتمال بسیار زیاد متفاوت از آدرس قبلی همان رایانه است. در نتیجه شناسایی کاربر از طریق آدرس آی پی ثبت شده در گزارش های فایل می تواند گمراه کننده باشد. با این حال، تحلیل بر اساس آی پی خود می تواند به عنوان یک روش مستقل برای مطالعه استفاده از وب به کار رود (برای نمونه نگاه کنید به دیویس، ۲۰۰۳)

گاه محققین سعی می کنند تا به جای شناسایی انفرادی کاربران در تحلیل گزارش وب به دسته بندی آنها بپردازند. از این روش در تحلیل گزارش وب سایت روزنامه های The Times/ The Sunday Times استفاده شد. (نیکلاس و دیگران، ۲۰۰۰). به این صورت که بر اساس حوزه آدرس یو.آر.ال<sup>۲۰</sup> (نظیر .org .co.uk .com) و آدرس آی پی مشخص شد که چه میزان از استفاده کنندگان از نهادهای دولتی (.org)، چه میزان از دانشگاه ها (مثلاً ut.ac.ir دانشگاه تهران)، چه میزان از موسسات تجاری و بخش خصوصی (.com) و یا بر اساس حوزه جغرافیای چه میزان از یک کشور بخصوص (مثلاً .uk بریتانیا) بوده اند. در این مورد نیز مشکلاتی وجود دارد به عنوان مثال بسیاری از آی.اس.پی ها در کشورهای مختلف آدرسهای آی پی خود را در آمریکا به ثبت می رسانند. در نتیجه در هنگام تحلیل، تمامی کاربرانی که از آن آی.اس.پی استفاده کرده اند به اشتباه آمریکایی تلقی می شوند حال آنکه ممکن است به عنوان مثال دریونان باشند.

<sup>15</sup> Robot

<sup>16</sup> Internet Crawler

<sup>17</sup> Internet Protocol: IP

<sup>18</sup> Arms

<sup>19</sup> Internet Service Provider: ISP

<sup>20</sup> Uniform Resource Locator: URL

نمونه این اشتباه در مطالعه روزنامه های فوق الذکر رخ داده است. بر اساس آدرس های آی.پی، حدود ۶۰ درصد خوانندگان وب سایت این روزنامه ها از آمریکا بوده اند حال اینکه بر اساس بانک اطلاعاتی مشترکین که نزد ناشر قرار دارد (خوانندگان این روزنامه ها جهت مطالعه شکل پیوسته آنها نیاز به ثبت مشخصات خود دارند)، تنها ۲۶ درصد از خوانندگان، خود را مقیم آمریکا معرفی کرده اند (نیکلاس و دیگران، ۲۰۰۰ ص ۴۰۷). {پایان صفحه ۱۸۱}

در مورد سایت هایی که از شناسه کاربری و کلمه عبور یا فناوری کوکی<sup>۲۱</sup> برای شناسایی کاربران تکراری و یا صدور مجوز دسترسی به محتوای خود استفاده می کنند، شناسایی کاربر قدری مطمئن تر است. با این حال در این مورد نیز وجود سرورهای پروکسی و دیواره های آتش<sup>۲۲</sup> که جهت افزایش کارایی و امنیت شبکه ها استفاده می شوند مانع از شناسایی دقیق و مطمئن کاربر می شود. این دو فناوری در حقیقت آدرس های آی.پی حقیقی کاربران را پنهان می کنند. بنا بر این شناسایی دقیق کاربر زمانی به نحو نسبتاً مطلوب میسر می شود که کاربر با استفاده از شناسه کاربری و رمز عبور و با اختصاص یک آدرس آی.پی ثابت و بدون وجود پروکسی و دیواره آتش از یک رایانه غیر عمومی به استفاده از وب بپردازد.

ه. طول زمان یک نشست<sup>۲۳</sup>. یکی از متغیرهای مورد استفاده در روش تحلیل گزارش های وب، تداوم یا طول زمان یک نشست است. تمام فعالیتها (کلیک های)ی یک کاربر از یک وب سایت بخصوص طی یک فاصله خاص زمانی یک نشست تلقی می شود (زاویتز، ۱۹۹۸). از آنجا کاربران معمولاً پس از استفاده از یک وب سایت خروج خود را رسماً ثبت نمی کنند (log off نمی کنند)، در نتیجه سنجش این متغیر معمولاً سؤال برانگیز است. برای رفع این نقیصه معمولاً به صورت پیش فرض زمان مشخصی به عنوان حداکثر زمان مجاز برای غیرفعال بودن کاربر در نظر گرفته می شود و پس از آن فعالیت کاربر به عنوان یک کاربر جدید یا نشست جدید در نظر گرفته می شود. به این معنی که اگر کاربری به مدت فرضاً ۳۰ دقیقه غیر فعال باشد (روی هیچ فرایبندی کلیک نکند)، سیستم چنین در نظر می گیرد که کاربر در حال استفاده از وب نیست.

### نرم افزارهای تحلیل گزارش های وب

همانطور که نویسنده مقاله متذکر شده (ص ۷۴) نرم افزارهای تجاری متعددی برای تحلیل گزارش های وب وجود دارد. اما نکته این است که اکثر این نرم افزارها به منظور گزارش گیری برای صاحبان وب طراحی شده اند و چندان جوابگوی نیازهای متخصصین اطلاع رسانی نیستند. در حقیقت این نرم افزارها بعضاً متغیرها و ملاک های از پیش تعریف شده را به پژوهشگر تحمیل می کنند. گروه سایبر به جهت نامناسب بودن این نرم افزارها از شیوه تجزیه داده ها<sup>۲۴</sup> و نرم افزار اس.پی.اس. اس<sup>۲۵</sup> جهت استخراج آمار استفاده می کند. این شیوه اگر چه زمان، منابع و تخصص بیشتری می طلبد اما از دقت بیشتری {پایان صفحه ۱۸۲} برخوردار است و بسته به اهداف تحقیق می تواند جوابگوی محقق باشد.

### ارتقاء کیفی روش تحلیل گزارش های وب

صرف نظر از اینکه می توان در کنار روش تحلیل گزارش وب از ترکیبی از روش های مختلف از جمله مصاحبه، پرسشنامه و مشاهده به صورت توأم جهت دستیابی به نتایج مطمئن تر استفاده کرد، با استفاده از فنون وترفندهای خاص می توان کیفیت خود روش تحلیل گزارش وب را نیز ارتقا بخشید.

یکی از فنون بکار گرفته شده توسط گروه سایبر روش تحلیل خرد<sup>۲۶</sup> است (نیکلاس و هانتینگتون، ۲۰۰۳، ص ۳۹۲). منظور از تحلیل خرد، تحلیل بخش مطمئنی از داده های گزارش وب (نه تمام داده ها) و ردگیری تعداد نسبتاً کمی از کاربران است. به صورت خاص می توان از سه شیوه برای تحلیل خرد استفاده کرد.

۱. جداسازی یک گروه خاص از کاربران که می توان در مورد منشأ جغرافیایی آنها مطمئن بود. این شیوه به عنوان مثال در مورد کاربران بریتانیایی که حوزه آدرس یو.آر.آل آنها با "uk" کاملاً مشخص است کاربرد دارد.
۲. تحلیل یک گروه جزئی از کاربران که احتمال اینکه آدرس های آی.پی آنها منعکس کننده استفاده افراد مشخصی باشد بیشتر است. به این معنی که آدرس های آی.پی از نوع ثابت بوده و مشخص هستند.
۳. ردگیری و بررسی استفاده های صورت گرفته توسط افراد خاص و نه گروه ها. در تحلیل گزارش های وب می توان داده های مربوط به یک کاربر منفرد را بر اساس اطلاعات مربوط به آدرس آی.پی، مشخصات مرورگر و شناسه کاربری جدا کرد و به صورت موردی بررسی نمود. (برای نمونه نگاه کنید به نیکلاس و هانتینگتون، ۲۰۰۳).

<sup>21</sup> Cooky

<sup>22</sup> Firewall

<sup>23</sup> Session

<sup>24</sup> Parsing

<sup>25</sup> Statistical Package for Social Sciences: SPSS

<sup>26</sup> Microanalysis

تدابیر دیگری نیز می‌توان بر حسب مورد تحقیق و ویژگی‌های سایت مورد بررسی بکار برد از جمله درموردی که اطلاعات مربوط به کاربران در یک بانک اطلاعاتی موجود است (نظیر مشترکین یک نشریه یا روزنامه الکترونیک) می‌توان آنها را به صورت اطلاعات تکمیلی برای شناخت کاربران و نیز سنجش میزان صحت نتایج حاصل از تحلیل به کار برد.

باید توجه داشت که محتوا و ساختار داده‌های ذخیره شده در گزارش‌های وب بسته به نوع سایت، نرم افزار به کار رفته برای ثبت داده‌ها و یا حتی سیاست صاحبان سایت‌ها {پایان صفحه ۱۸۳} تفاوت می‌کند. به عنوان مثال ساختار و نوع داده‌ها در فایل گزارش سایت یک ناشر نشریات الکترونیکی با گزارش‌های یک وب سایت عمومی تفاوت زیادی دارد (برای نمونه نگاه کنید به نیکلاس و دیگران، ۲۰۰۱؛ ۲۰۰۲؛ ۲۰۰۳). همانطور که مشکلات ذکر شده در این نوشته را نمی‌توان در مورد تمام انواع فایل‌های گزارش وب تعمیم داد، در هنگام بکارگیری شیوه تحلیل نیز باید به ویژگی‌های خاص آن سایت و گزارش‌های آن اندیشید.

#### خاتمه

در خاتمه ذکر این نکته نیز لازم است که اگر چه روش تحلیل گزارش‌های وب افق‌های تازه‌ای در مطالعه خیل کثیرکاربران اطلاعات الکترونیکی گشوده و تکامل این روش چه در مرحله ثبت داده‌ها در فایل‌های گزارش و چه در مرحله تحلیل همچنان ادامه دارد، اما شیوه منطقی آن است که همواره با احتیاط به نتایج این گونه پژوهش‌ها نگریست. چه بسا در مواردی استفاده از این روش برای طرح سؤال مناسب‌تر است تا رسیدن به پاسخ سئوالاتی خاص. مطالعه رفتار کاربر با استفاده از این روش سئوالات زیادی در مورد رفتار آنها برمی‌انگیزد که پاسخ آنها را باید با شیوه‌های کیفی جست. هنگام استفاده از این روش باید به خاطر داشت که وب در حقیقت یک نظام برای توزیع صفحات وب است و نه یک نظام ذخیره و بازیابی اطلاعات. آنچه که در داده‌های فایل گزارش وب نهفته است بیانگر چیزی است که کاربران (با واسطه یک رایانه) انجام داده‌اند، نه آنچه که فکر می‌کنند انجام داده‌اند یا تصور می‌کنند بایستی انجام می‌دادند.

#### منابع

Arms, W. Y. (2000). *Digital Libraries*. Cambridge: MIT press.

Davis, P. M., and Solla, L. R. (2003). An IP-Level analysis of usage statistics for electronic journals in chemistry: making inferences about user behavior. *Journal of the American Society for Information Science and Technology* 54(11), 1062-1068.

Fieber, John, Browser caching and Web log analysis, [Online] available at:

<http://ella.slis.indiana.edu/~jfieber/papers/bcwla/bcwla.html>

last visited 18/02/2004

{پایان صفحه ۱۸۴}

Gutzman, A. (1999). Analysing traffic on your E-commerce site [Online] available at:

[http://ecommerce.internet.com/news/insights/ectech/article/0,,9561\\_00.html](http://ecommerce.internet.com/news/insights/ectech/article/0,,9561_00.html)

last visited 18/02/2004

Nicholas, D., and Huntington, P. (2003). Micro-mining and segmented log file analysis: a method for enriching the data yield from Internet log files. *Journal of Information Science* 29(5), 391-404.

Nicholas, D., Huntington, P., Lievesley, N., and Wasti, A. (2000). Evaluating consumer website logs: a case study of The Times/The Sunday Times website. *Journal of Information Science* 26(6), 399-411.

Nicholas, D., Huntington, P., Lievesley, N., and Withey, R. (1999). Cracking the code: Web log analysis. *Online & CD-ROM Review* 23(5), 263-265.

Nicholas, D., Huntington, P., and Watkinson, Anthony. (2003). Digital journals, Big Deal and online searching behaviour: a pilot study. *Aslib Proceedings: New Information Perspectives*. 55(1-2): 84-109.

Nicholas, D., Huntington, P., and Williams, P. (2001). Establishing metrics for the evaluation of touch screen kiosks *Journal of Information Science*. 27(2): 61-71

Nicholas, D., Huntington, P., and Williams, P. (2002). Evaluating metrics for comparing the use of Web sites: a case study of two consumer health web sites. *Journal of Information Science* 28(1), 63-75.

Zawitz, M. W. (1998) Web statistics- measuring user activity. [Online] available at:  
<http://www.ojp.usdoj.gov/bjs/pub/ascii/wsmua.txt>  
last visited 18/02/2004

{پایان صفحه ۱۸۵}