دانشگاه جامع امام حسین(ع)
دانشکده و پژوهشکده فنی و مهندسی

مجله علمی-پژوهشی «علوم و فناوری‌های پدافند غیرعامل»
سال اول، شماره ۲، زمستان ۱۳۸۹: ص ۴۳-۳۳

Available online:
www.adst.ir

تحلیل رفتار کاربران برای بهبود پالایش دسترسی به سایت‌ها

مهدی نقوی[1*]، وحید نقوی[2]، بهروز مینایی[3]

۱- دانشجوی دکتری، ۳- استادیار، دانشگاه علم و صنعت ایران، دانشکده کامپیوتر

۲- دانشجوی کارشناسی ارشد، دانشگاه آزاد اسلامی، واحد اراک

E-mail: naghavi@iust.ac.ir

(دریافت: ۸۸/۰۹/۰٤ ، پذیرش: ۸۹/۰٦/۲۹)

چکیده

در این مقاله روش نوینی جهت بهبود پالایش دسترسی به سایت‌ها ارائه شده است. این روش مبتنی بر تحلیل رفتار بازدیدکنندگان صفحات وب می‌باشد. برای تحلیل رفتار کاربران، معماریی ارائه شده است که در سه مرحله بر اساس اطلاعات موجود در فایل‌های گزارش سرورهای پروکسی، الگوی رفتار کاربران وب براساس یک موتور داده‌کاو، استخراج شده و در بانک اطلاعات ذخیره می‌گردد. هنگام درخواست کاربر برای یک صفحه جدید، با بررسی و تحلیل سابقه رفتار کاربر، دسترسی مجاز و یا غیر مجاز به صفحه مورد نظر، اعلام می‌گردد. معماری ارائه شده با استفاده از داده‌های یکی از سایت‌های معتبر جهانی آزمایش شده و نتایج آن مورد بررسی و تحلیل قرار گرفته است.

کلیدواژه‌ها: پالایش دسترسی به وب؛ تحلیل رفتار؛ فیلترینگ؛ کاوش وب؛ مسدود کردن سایت‌ها

Analyzing Web Users Behavior to Improve Web Filtering

M. Naghavi[1*], V. Naghavi[2], B. Minaei[3]
Department of Computer, Iran University of Science and Technology
E-mail: naghavi@iust.ac.ir

Abstract

In this paper, a new method to improve sites access filtering is presented. This method is based on web visitor behavior analysis. In order to analyze the behavior of users, A procedure is proposed following which, based on the information in the log files of proxy servers, user's behavior history was extracted and stocked in a data bank by a data-mining engine. Authorized or unauthorized access to the desired page, when a user requests a new page, is reported after investigation of the user's behavior history. The above architecture was tested, using data from one of the globally credible sites, and the results were analyzed.

Keywords: Behavior Analysis; Web Mining; Web Usage Mining; Web Filtering; Web Blocking

۱. مقدمه

گسترش روزافزون دنیای وب باعث بوجود آمدن نیازهای جدیدی شده است که یکی از آن‌ها، پالایش سایت‌ها بر اساس اهداف مختلف است. طبق گزارش WebSpy در سال ۲۰۰۸، روزانه حدود ۴۰ میلیون کاربر فقط به منظور سرگرمی و گذراندن وقت وارد شبکه اینترنت می‌شوند[۱]. یک سوم وقت کارمندان شرکت‌های آمریکایی در امور غیر مرتبط با کار در اینترنت صرف شده و هزینه‌های ناشی از آن بالغ بر ۸۵ میلیارد دلار در سال برآورد شده است[۲]. طبق گزارش Covenant Eyes[1] از هر ۵ نفر، ۲ نفر جهت بازدید از سایت‌های غیر اخلاقی به اینترنت متصل می‌شوند. کلمه sex دارای رتبه اول کلمات جستجو در موتورهای جستجو بوده و سایر کلمات مرتبط نیز از رتبه‌های بالا برخوردار می‌باشند[۳].

این مقاله با رویکرد تحلیل رفتار بازدیدکنندگان وب، نیاز فوق را مورد توجه قرار داده و در راستای بهبود فیلترینگ سایت‌ها ارائه شده است. در ابتدا موضوع پالایش دسترسی به وب مورد بررسی قرار گرفته و سپس مبانی مربوط به کاوش رفتار در وب بیان شده است. در ادامه، کارهای مرتبط با موضوع کاوش رفتار در وب توضیح داده شده و به بیان طرح مسئله و ارائه راه‌حل پرداخته و در نهایت چگونگی معماری استخراج الگو و تحلیل رفتار کاربر ارائه شده است. در بخش‌های بعدی به عنوان نمونه به بررسی آزمایشات صورت گرفته بر روی داده‌های پایگاه AOL پرداخته شده و در نهایت چالش‌های پیش‌رو و پیشنهاد کارهای آینده بیان شده است.

۲. پالایش دسترسی به وب

عوامل فوق، سازمان‌ها و شرکت‌ها را بر آن داشت که جهت کم کردن سطح استفاده نادرست از اینترنت، با استفاده از روش‌های فیلترینگ، سایت‌های غیرمرتبط با موضوعات کاری و یا اهداف از پیش‌تعریف شده را مسدود نمایند. بسیاری از فیلترهای وب از الگوریتم سادهٔ کنترل وجود کلمات کلیدی[2] ممنوعه در کلمات تشکیل دهنده درخواست کاربر و URL (Uniform Resource Locator) استفاده می‌کنند. اگر چه این روش دارای مزایایی از جمله سادگی و سرعت مناسب می‌باشد، ولی قدرت تشخیص مفهوم یک کلمه در جمله را ندارد و همین امر سبب می‌شود تا تعداد زیادی از صفحات مجاز مسدود شوند. به طور مثال کلمه "essex" که نام یک دانشگاه و چندین شهر در نقاط مختلف دنیا است، به سبب داشتن حروف کلمه sex، ممنوعه منظور شده و فیلتر، آن را غیرمجاز تشخیص داده و صفحه مربوطه را مسدود می‌کند و یا ممکن است کاربر با یک هدف معتبر از چند کلمه مربوطه برای جستجوی موضوع مورد نظر خود استفاده کند. به طور مثال بسیار طبیعی است که یک پزشک برای تحقیق در موضوع تاثیر جنسیت در بیماری‌های قلبی از واژه sex در جستجوهایش استفاده نماید، درصورتیکه فیلتر، آن را غیرمجاز تشخیص داده و صفحه مربوطه را مسدود می‌نماید. از طرف دیگر ممکن است به دلیل عدم وجود کلمات کلیدی ممنوعه در درخواست کاربر و URL مربوط به یک صفحه غیر مجاز، فیلتر این صفحه را مجاز شناخته و اجازه عبور دهد.

در این مقاله روشی ارائه شده است که با تحلیل رفتار بازدیدکنندگان وب و با استفاده از تکنیک‌های کاوش در وب، پالایش دسترسی به وب را بهبود می‌بخشد. با توجه به موارد فوق، روش ارائه شده، برای فیلترهایی‌کارایی‌خواهد داشت که از روش کنترل کلمات ممنوعه استفاده می‌کنند.

۳. کاوش رفتار در وب

رشد روزافزون صفحات وب باعث پدید آمدن حجم وسیعی از اطلاعات غیر ساخت‌یافته در شبکه اینترنت شده است[۴] که تا آوریل ۲۰۰۸ بیش از ۱۶۵/۷ میلیون سایت وب به صورت فعال مشغول به کار شده‌اند[۵]. در چنین وضعیتی کاوش هدفمند در وب به‌عنوان یک مسئله اساسی در دنیای وب مطرح و به یکی از محورهای تحقیقاتی تبدیل شده است.

کاوش در وب[3] را می‌توان به سه محور کاوش محتوای وب[4]، کاوش ساختار وب و کاوش رفتار در وب[5] دسته‌بندی کرد[۶]. در این مقاله بر روش کاوش رفتار در وب تمرکز شده و در ادامه به طور مختصر تشریح می‌گردد.

در تکنیک کاوش رفتار در وب با استفاده از داده‌کاوی در

---

۱. یکی از شرکت‌های نرم‌افزارهای فیلترینگ که اغلب فعالیت‌های اینترنتی را گزارش می‌کند.

۲. کلمه کلیدی به کلماتی می‌گوییم که وجود آن‌ها در درخواست کاربر و URL احتمالا نشانگر سمت و سوی کاربر به یک موضوع ممنوعه (غیر مرتبط با اهداف سازمان‌ها و شرکت‌های مربوطه) می‌باشد.

3. Web Mining
4. Web Content Mining
5. Web Usage Mining

فایل‌های گزارش[1] سرورهای پروکسی و یا میزبان، سعی در درک رفتار کاربر و پیش‌بینی آن در آینده می‌شود. تحلیل اطلاعات فایل‌های گزارش و داده‌های ثبت شده کاربران، می‌تواند اطلاعات مهم و قابل توجهی برای صاحبان سایت‌ها و سازمان‌های مربوطه فراهم کند.

۳-۱. ساختار فایل گزارش

جدول (۱) نمونه‌ای از فایل گزارش سرورهای پروکسی را نشان می‌دهد[۷].

هر سطر فایل گزارش، مربوط به دسترسی یک کاربر می‌باشد. مشخصه‌های معمول این فایل در جدول (۲) آمده است که برای آگاهی بیشتر از توضیحات این مشخصه‌ها می‌توان به [۵] رجوع نمود.

۳-۲. مراحل کاوش رفتار در وب

کاوش رفتار در وب شامل سه مرحله اصلی است. مرحله اول، آماده‌سازی داده‌ها است. در این مرحله داده‌های فایل گزارش از اطلاعات غیر ضروری و غیر مفید پاکسازی شده و اطلاعات مورد نیاز به صورت یکپارچه جمع‌آوری و در پایگاه داده مناسبی ذخیره می‌گردد[۸].

مرحله دوم که اصلی‌ترین مرحله کاوش رفتار در وب می‌باشد، استخراج الگوی رفتار کاربر است[۹]. در این مرحله براساس اطلاعات دسترسی کاربر و با توجه به اهداف مربوط به تحلیل، اطلاعات مورد نظر به‌طور دقیق مورد بررسی قرار گرفته و الگوی رفتاری کاربر استخراج می‌شود. از مشکل‌ترین قسمت‌های این مرحله شناخت نشست کاربر است. معمولاً اتصال کاربران به سایت‌های عمومی بدون نام و مشخصات است. همچنین نمی‌توان به آدرس IP او نیز متکی بود. زیرا بخشی از کاربران به واسطه ISPها و یا ICPها به اینترنت متصل شده و ممکن است آدرس IP این واسطه‌ها به جای آدرس مبدأ ارسال شده باشد. در این صورت جهت شناخت کاربر باید از

جدول ۱. نمونه‌ای از فایل گزارش دسترسی

```
2006-02-01 00:08:43 1.2.3.4 - GET /classes/cs589/papers.html - 200 9221
HTTP/1.1 maya.cs.depaul.edu
Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1;+.NET+CLR+2.0.50727)
http://dataminingresources.blogspot.com/
2006-02-01 00:08:46 1.2.3.4 - GET /classes/cs589/papers/cms-tai.pdf - 200 4096
HTTP/1.1 maya.cs.depaul.edu
Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1;+.NET+CLR+2.0.50727)
http://maya.cs.depaul.edu/~classes/cs589/papers.html
2006-02-01 08:01:28 2.3.4.5 - GET /classes/ds575/papers/hyperlink.pdf - 200
318814 HTTP/1.1 maya.cs.depaul.edu
Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1)
http://www.google.com/search?hl=en&lr=&q=hyperlink+analysis+for+the+web+survey
2006-02-02 19:34:45 3.4.5.6 - GET /classes/cs480/announce.html - 200 3794
HTTP/1.1 maya.cs.depaul.edu
Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1)
http://maya.cs.depaul.edu/~classes/cs480/
2006-02-02 19:34:45 3.4.5.6 - GET /classes/cs480/styles2.css - 200 1636
HTTP/1.1 maya.cs.depaul.edu
Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1)
http://maya.cs.depaul.edu/~classes/cs480/announce.html
2006-02-02 19:34:45 3.4.5.6 - GET /classes/cs480/header.gif - 200 6027
HTTP/1.1 maya.cs.depaul.edu
Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1)
http://maya.cs.depaul.edu/~classes/cs480/announce.html
```

1. Server Access Log

تکنیک‌های دیگری مانند دسته‌بندی نشست[1]های کاربران استفاده کرد[۱۰].

مرحله سوم، تحلیل الگوهای استخراج شده است. در این مرحله الگوهای استخراج شده، مورد بازبینی قرار گرفته و رفتار کاربر مورد تحلیل قرار می‌گیرد. به‌عنوان مثال درمی‌یابیم که کاربر پیگیر موسیقی بوده و همچنین به چه نوع موسیقی علاقه‌مند است[۱۱].

جدول ۲. مشخصه‌های اصلی فایل گزارش دسترسی

Remote Host	ng21.exabot.com
Login Name	-
Authorised User	-
Date / Time	[01/Oct/2004:22:27:46 +0100]
Request	"Get /staff/asga/context_learn HTTP/1.1"
Status	301
Bytes Sent	388

۴. کارهای انجام شده مرتبط

گانزن باتامالای، مقاله وب‌سایت‌های خودسازگار را بر اساس تحقیقات مربوط به پایان‌نامه کارشناسی ارشد خود در دانشگاه رابرت گاردن[2] را در سال ۲۰۰۸ ارائه کرده است[۵]. او با استفاده از کاوش گزارش‌های دسترسی، الگوی رفتار کاربر را استخراج کرده و بر اساس آن سعی در یافتن سریع‌تر صفحات هدف مورد نظر کاربر کرده است.

در این مقاله روشی ارائه شده که راه‌های کوتاه‌تری برای دسترسی به صفحات هدف شناسایی کرده و ضمن پاکسازی فایل گزارش از اطلاعات ناخواسته، جلسه‌بندی[3] کاربران صورت می‌گیرد. منظور از جلسه‌بندی، شناسایی و جداسازی آن‌دسته از صفحاتی است که کاربر در یک نشست در طی اتصال به وب‌سایت، به آن صفحات دسترسی می‌یابد.

در این مقاله، زمان متوسط صرف شده یک کاربر بر روی یک صفحه حدود ۴۰ ثانیه محاسبه شده و بر اساس نمودارهای ارائه شده، زمان متوسط یک نشست ۳ دقیقه برآورد شده است. این زمان‌ها برای بدست آوردن مجموعه صفحات مرتبط به هم در یک نشست کاربر لازم است.

جهت شناسایی و یافتن صفحه هدف کاربر، از روش‌های مدرک پایانی[4] و زمان دیدن[5] استفاده شده است که در این روش، صفحه پایانی فراوانی صفحات مورد دستیابی، محاسبه و در یک جدول قرار می‌گیرد. جدول (۳) یک نمونه از میزان فراوانی صفحات را نشان می‌دهد. عامل EDF[6] تعداد مراجعات به صفحه مربوطه در آخرین نشست کاربر را نشان می‌دهد.

جدول ۳. محاسبه فراوانی صفحات پایانی

ID.P	EDF	Page Name
P۱	۱۲	/staff/dc/hcow/menu.html
P۲۴	۲۷	/welcome.ht
P۲۹	۵	/courses/Undergraduate/elective.htm
P۳۶	۱۸	/staff/dc/hcow/links.html
P۳۹	۲۹	/courses/postgradindex.htm
P۵۹	۴	/staff/khui/teaching/cm3016/
…	…	…

عامل دیگری به نام IV[7] تعریف شده که تعداد دفعاتی که یک صفحه دیده شده است را نشان می‌دهد. با تعریف این عامل، شرط $EDF/IV \geq$ برای پذیرفتن صفحات، تعیین شده است. این شرط اجبار می‌کند که حداقل نیمی از صفحات مورد دستیابی در آخرین نشست، مربوط به صفحه پایانی باشد که در جدول (۴)، مقادیر عامل‌ها و نتیجه شرط مشخص شده است.

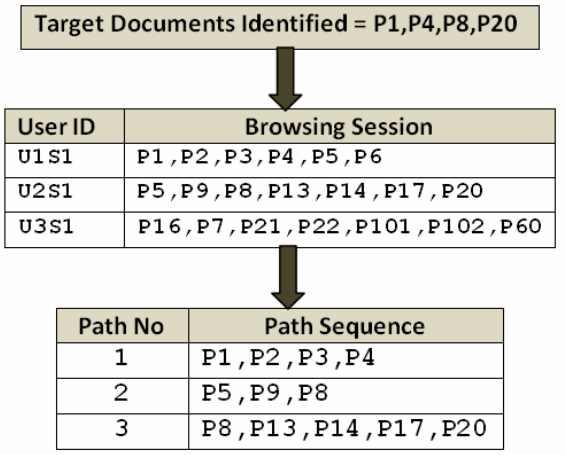
با توجه به جدول (۴) صفحات P36 و P59 به عنوان صفحات هدف گزینش می‌شوند.

جدول ۴. محاسبه نتایج شروط

ID.P	EDF	IV	EDF/IV	Con.
P۱	۱۲	۱۰۲	۰/۱۲	No
P۲۴	۲۷	۱۳۸	۰/۰۲	No
P۲۹	۵	۱۱	۰/۰۵	No
P۳۶	۱۸	۳۱	۰/۵۸	Yes
P۳۹	۲۹	۶۷	۰/۴۳	No
P۵۹	۴	۸	۰/۵۰	Yes
…	…	…	…	…

پس از آنکه صفحات هدف مشخص شدند، باید مسیرها استخراج شوند. منظور از مسیرها، استخراج الگوی دسترسی به صفحات است.

1. Session Clustering
2. Robert Gordon University
3. Sessionization
4. End Document Method
5. Browsing Time Method
6. End Document Frequency
7. Incoming Visits



شکل ۱. مثالی از استخراج مسیرها

برای این منظور از میان صفحات دسترسی شده توسط هر کاربر، مسیرهایی انتخاب می‌شوند که ابتدا و انتهای آن‌ها جزء صفحات هدف باشند. شکل (۱) مثالی از انتخاب مسیرها است.

در نهایت با استفاده از الگوریتم k-means مسیرها دسته‌بندی می‌شوند. سپس با استفاده از آن، گره‌هایی برای دسترسی‌های سریع استخراج می‌شود که شکل (۲) مثال‌هایی از ایجاد این نقطه‌های دسترسی‌های سریع را نشان می‌دهد.

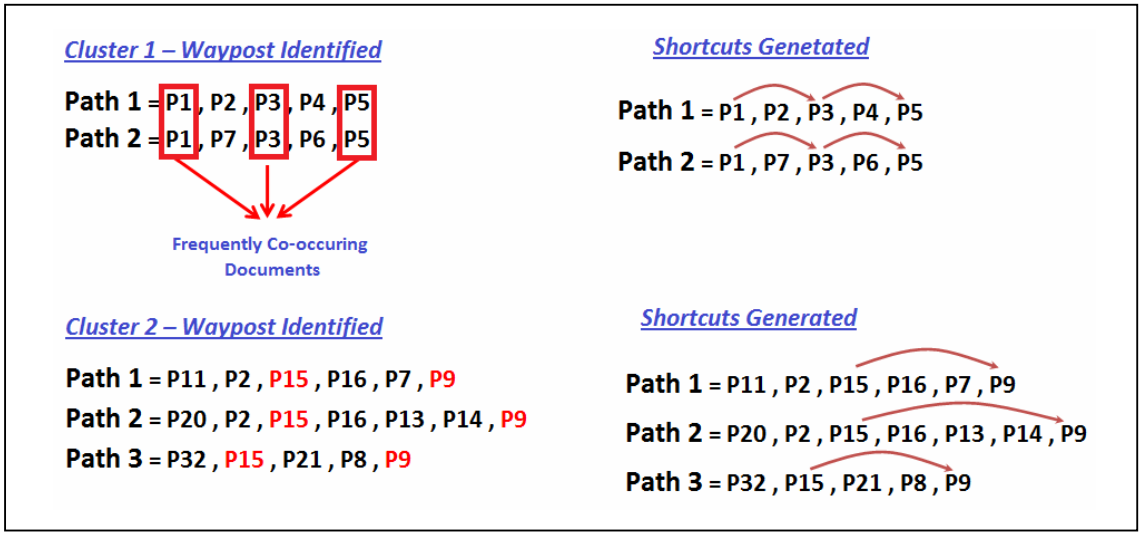
همان‌طور که در شکل (۲) دیده می‌شود، صفحات P۱، P۳ و P۵ در مثال اول و صفحات P۱۵ و P۹ در مثال دوم به عنوان نقاط دسترسی سریع منظور شده و ایجاد می‌شوند.

۵. طرح مسئله و ارائه راه‌حل

پالایش صفحات وب بر اساس کلمه‌های کلیدی ممنوعه، یک راه حل افراطی بوده و باعث پایین آمدن کارایی و نارضایتی کاربران می‌گردد. ممکن است هدف کاربران از به‌کار بردن کلمات کلیدی ممنوعه، ورود به صفحات ممنوعه نباشد، ولی چون ملاک ما فقط کلمات ممنوعه است، به صورت افراطی از دسترسی کاربر به صفحات مورد نظرش جلوگیری به عمل می‌آید. اگر بتوان سابقه رفتاری کاربر را استخراج کرده و با بررسی درخواست جاری وی، قصد او را از این درخواست تخمین زده و عکس‌العمل مناسب‌تری نشان دهیم، توانسته‌ایم یک روش عادلانه‌تری را ارائه دهیم. به همین منظوردر ادامه معماری تحلیل رفتار کاربر ارائه شده و تشریح می‌گردد.

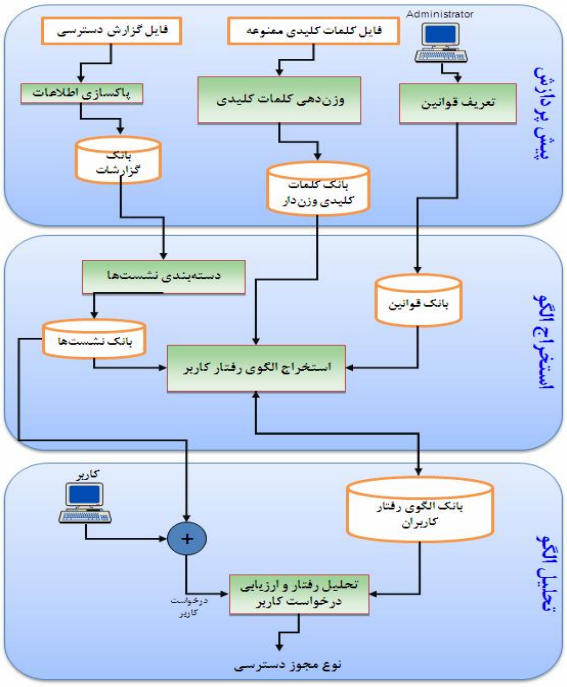
۶. معماری تحلیل رفتار کاربر

شکل (۳)، معماری پالایش درخواست‌های کاربر، بر اساس رفتار او را نشان می‌دهد. بر اساس این معماری، در سه مرحله عمل پالایش صورت می‌پذیرد:



شکل ۲. مثالی از استخراج نقاط دسترسی سریع

مرحله اول پیش پردازش است که در این مرحله باید سه پردازش اساسی صورت بگیرد؛ فایل گزارش[1] باید به صورت یک بانک اطلاعات آماده‌سازی و قابل استفاده گردد. به همین منظور باید داده‌ها را از داده‌های غیر مفید پاکسازی نمود. در این پالایش ضمن تبدیل اطلاعات قبلی به شکل ساخت‌یافته، فیلدهای مورد نیاز ذخیره شده و فیلدهای غیر ضروری دور ریخته می‌شود.



شکل ۳. معماری پالایش درخواست‌ها بر اساس رفتار کاربر

هنگامی که کاربر یک صفحه خاص را دریافت می‌کند، ممکن است در فایل گزارش سرور چندین سطر گزارش مربوط به این درخواست ثبت شود و در بیشتر مواقع ممکن است فقط یک سطر این اطلاعات مورد نیاز باشد که در این‌صورت باید این اطلاعات اضافه نیز حذف گردد. به‌عنوان مثال اطلاعات مربوط به فایل‌های غیر مرتبط مانند فایل‌های تصویری می‌تواند حذف گردد.

در همین مرحله باید به کلمه‌های کلیدی وزن خاصی را نسبت داده و در پایگاه داده مربوطه نگهداری نمود. این وزن‌دهی توسط مدیر سیستم فیلترینگ انجام می‌شود و باید متناسب کاربرد آن‌ها در جستجوهای ممنوعه، صورت پذیرد. وزن‌ها در جدولی مانند جدول (۵) ذخیره می‌شوند.

جدول ۵. نمونه‌ای از جدول وزن‌دهی کلمات کلیدی

ردیف	کلمه	وزن
۱	sensation	۳۰
۲	bosom	۳۰
۳	Sex	۷۰
۴	Bare	۷۰
۵	couple	۵۰
۶	plump	۵۰
۷	whore	۹۰
۸	girl	۳۰
۹	porn	۸۰
۱۰	adult	۶۰

همچنین در این مرحله باید توسط مدیر سیستم، قوانین مورد نیاز در بانک قوانین ثبت شود. این بانک، قوانینی مانند قانون وزن‌دهی به کاربران در گروه‌های مختلف و یا قوانین مربوط به تعیین مجوز دستیابی با توجه به رفتار کاربر را در خود ذخیره می‌نماید. از دیگر قوانینی که می‌توان در بانک قوانین ثبت کرد، اطلاعات مربوط به مرزبندی گروه‌های بازدیدکننده وب و نحوه تعیین آن‌ها خواهد بود که در بخش بعدی توضیح داده می‌شوند.

مرحله دوم، مرحله استخراج الگوی رفتار است. در این مرحله باید برای شناسایی کاربران و انتساب کد شناسایی، نشست‌های ایشان را دسته‌بندی کرده و در پایگاه داده مربوطه ذخیره نمود. بر اساس اطلاعات موجود در این پایگاه و پایگاه‌های کلمه‌های کلیدی وزن‌دار، قوانین و همچنین بر اساس سابقه رفتار قبلی کاربر، رفتار وی مورد بررسی قرار گرفته و پس از استخراج الگوی رفتاری، در بانک مربوطه ذخیره می‌گردد. چگونگی استحراج الگوی رفتار کاربران در بخش ۳ توضیح داده شده است.

در مرحله تحلیل الگو، پس از استخراج الگوی رفتاری کاربر، درخواست‌های او بر اساس رفتار گذشته ارزیابی شده و اجازه دسترسی به صفحه مورد درخواست مشخص می‌شود.

۷. تحلیل رفتار بازدیدکنندگان

هدف این بخش تبیین چگونگی تحلیل رفتار بازدیدکنندگان

1. Log File

وب بر اساس اطلاعات گزارش دسترسی کاربران می‌باشد. جهت بهبود فیلتر کردن سایت‌ها از سابقه رفتاری هر کاربر کمک گرفته و معین می‌گردد که دسترسی جاری کاربر با چه هدفی صورت گرفته و مجاز و یا غیر مجاز بودن دسترسی اعلام می‌شود. به همین منظور کاربران بر اساس رفتار کاربران در مراجعه به صفحات وب در سه دسته‌بندی تعریف می‌گردند: گروه کاربران با رفتارهای هنجار، ناهنجار و نامشخص.

رفتارهای هنجار به آن دسته از رفتارهای کاربر گفته می‌شود که از دید سیستم فیلترینگ، جستجو و گشت و گذار عادی در وب باشد. اگر کاربر از روال عادی خارج شده و سعی در دسترسی به صفحات ممنوعه را داشته باشد به‌عنوان کاربر با رفتار ناهنجار تعریف می‌شود. به‌عنوان مثال کاربر به صفحات مربوط به معاملات مواد مخدر و یا در ساعات اداری به سایت‌های غیر مرتبط با کار متصل گردد.

گروه کاربران با رفتار نامشخص به کاربرانی گفته می‌شود که رفتار آنان به روشنی مشخص نشده و در حال بررسی است. این کاربران گاهی دارای رفتار عادی و گاهی نیز دارای رفتار غیر عادی بوده و نمی‌توانیم تشخیص دهیم این کاربران دارای رفتار هنجار و یا ناهنجار هستند.

۷-۱. گروه‌بندی کاربران

در ابتدای ورود کاربر به سیستم، برای او وزن رفتاری اولیه منظور شده به نحوی که در گروه هنجار دسته‌بندی شود. منظور از وزن رفتاری، عددی است بین صفر تا صد که نشانگر رفتار کاربر می‌باشد. هرچه این عدد به صد نزدیک‌تر باشد نشانگر رفتار ناهنجار بیشتر و هر چه به صفر نزدیک‌تر باشد، نشانگر ناهنجاری کمتر و در واقع رفتار سالم‌تر و هنجارتر می‌باشد.

در صورتی که کاربری از گروه هنجار صفحه ممنوعه‌ای را تقاضا بکند، آنگاه به او اجازه دیدن صفحه درخواستی داده می‌شود، ولی رفتار او به صورت دقیق‌تری مورد بررسی قرار می‌گیرد. در صورتی که کاربر، اصرار به ادامه رفتارهای ناهنجار را داشته باشد، وزن رفتاری او در گروه ناهنجار افزوده شده و دسترسی‌ها برای وی محدودتر می‌گردد.

۷-۲. الگوریتم اختصاص وزن رفتاری

با توجه به رفتار کاربر، یک عدد متناسب با حضور او در گروه‌های فوق نسبت داده می‌شود. این عدد که وزن رفتاری نامیده می‌شود، در تحلیل رفتار کاربر به ما کمک خواهد کرد. برای اختصاص وزن باید از قوانینی که قبلاً تعریف شده و در بانک قوانین موجود است، استفاده کنیم.

برای تعیین رفتار کاربر باید به چند نکته توجه کرد؛ نخست اینکه یک کاربر می‌تواند رفتار کاملاً هنجار و یا رفتار کاملاً ناهنجار داشته باشد (رفتار کاملاً هنجار را معادل وزن صفر در نظرگرفته و رفتار کاملاً ناهنجار معادل وزن صد تعریف می‌گردد).

دوم اینکه برای تعیین رفتار (W) یک فرد در زمان t باید رفتار گذشته (W) او را نیز لحاظ کنیم.

سوم اینکه هر چند سابقه رفتار فرد در تعیین رفتار جدید فرد تاثیر دارد، اما تغییر رفتار او نیز باید بتواند به سرعت وضعیت رفتاری او را اصلاح نماید. باید توجه داشت که ما قصد نداریم کاربر را به لحاظ رفتاری ارزیابی نماییم، بلکه به‌دنبال این هستیم که رفتار زمان حال او را با توجه به گذشته‌اش بهتر درک کنیم. منظور از رفتار گذشته، گذشته نزدیک در همان نشست بوده و به رفتار کاربر در نشست‌های قبلی توجه نخواهیم کرد.

با توجه به موارد فوق، فرمول (۱) پیشنهادی برای محاسبه وزن رفتاری کاربر در زمان t می‌باشد:

$$W(t) = a.W(t-1) + b.\sum_i^n \qquad (۱)$$

W(t) رفتار جاری کاربر و وزن کلمات کلیدی ممنوعه بوده که کاربر در درخواست جاری خود استفاده کرده و از جدول مربوط به وزن کلمات کلیدی استخراج می‌گردد و W(t-1) رفتار محاسبه شده او در آخرین دسترسی‌اش به صفحات وب می‌باشد. ضرایب a و b عددی بین ۰ و ۱ است. ضریب a میزان تأثیر رفتار گذشته فرد در محاسبه رفتار جاری او و ضریب b سهم تأثیر کلمات کلیدی به‌کار رفته در درخواست جاری او را در تعیین رفتار جاری فرد نشان می‌دهد. مقدار n نشان دهنده تعداد کلمات کلیدی منظور شده یک درخواست می‌باشد.

همان‌طور که قبلاً اشاره شد، مقدار W حداکثر ۱۰۰ می‌باشد و اگر مقدار محاسبه شده در مواردی بیش از عدد ۱۰۰ به‌دست آید، باید آن را ۱۰۰ در نظر گرفته و در محاسبات مرحله بعد لحاظ کرد.

به‌عنوان مثال اگر تعداد کلمات کلیدی در یک درخواست که مؤثر در تعیین رفتار جاری کاربر باشد را ۳ فرض کرده و ضرایب a و b را به ترتیب $\frac{3}{4}$ و $\frac{1}{6}$ منظور نماییم، فرمول فوق به صورت فرمول (۲) در می‌آید.

$$W(t)=\left(\frac{3}{4}W(t-1)\frac{1}{6}(W1+W2+W3)\right) \qquad (۲)$$

در فرمول (۲)، W نشانگر آخرین وزن به‌دست آمده برای هر کاربر در زمان t و در نشست جاری می‌باشد و فقط ۳ کلمه از کلمات کلیدی ممنوعه که کاربر به کار می‌برد محاسبه می‌شود. یعنی اگر تعداد کلمات کلیدی ممنوعه که کاربر به کار می‌برد بیش از سه کلمه باشد، فقط سه کلمه از آن انتخاب می‌شود. این سه کلمه ممنوعه می‌تواند سه کلمه اول ممنوعه بوده و یا آن‌کلماتی انتخاب شوند که بیشترین وزن را دارا می‌باشند.

در ابتدای کار، مقدار ۵۰ به عنوان وزن اولیه به رفتار کاربر اختصاص داده می‌شود. کاربری که در اولین و دومین درخواست او هیچ واژه ممنوعه‌ای وجود نداشته باشد، وزن او به صورت فرمول‌های (۳) و (۴) محاسبه می‌گردد.

$$W=TRUNC\left(\frac{3\times 50}{4}+\frac{0+0+0}{6}\right)=37 \qquad (۳)$$

$$W=TRUNC\left(\frac{3\times 37}{4}+\frac{0+0+0}{6}\right)=27 \qquad (۴)$$

و اگر در درخواست سوم او سه کلمه ممنوعه با وزن‌های ۹۰، ۶۰ و ۷۰ وجود داشته باشد، وزن او در قسمت کلمات ممنوعه ۵۶ محاسبه می‌شود. اگر در درخواست چهارم، کلمات ممنوعه با وزن‌های ۶۰، ۷۰ و ۸۰ و در درخواست پنجم، کلمات ممنوعه با وزن‌های ۹۰، ۹۰ و ۹۰ را به‌کار برند، وزن رفتاری کاربر ۷۷ محاسبه می‌گردد.

لازم به ذکر است در محاسبه آخر، مقدار بدست آمده بزرگ‌تر از ۱۰۰ بوده که باید این مقدار کوچک‌تر و یا مساوی ۱۰۰ شود.

همان‌طور که دیده می‌شود سعی بیشتر برای جستجوی کلمات ممنوعه و یا آدرس‌دهی آن‌ها، وزن رفتار کاربر را در گروه کلمات ممنوعه بالا می‌برد. اگر کاربر رفتار خود را اصلاح کرده و کلمات ممنوعه کمتری به‌کار برده و یا اصلا به کار نبرد، وزن او نیز رو به کاهش می‌رود. به‌طور مثال اگر در دو درخواست بعدی فقط یک کلمه ممنوعه با وزن ۳۰ وجود داشته باشد، وزن او ابتدا به ۸۰ و سپس به ۶۵ کاهش می‌یابد.

۸. استفاده از تحلیل رفتار برای پالایش دسترسی کاربران به وب

همان‌طور که در فصل قبل بیان گردید، می‌توان با استفاده از فایل‌های گزارش دسترسی کاربران به صفحات مختلف، رفتار آنان را مورد ارزیابی قرار داده و جهت استفاده در تصمیم‌گیری‌های بعدی نگهداری کرد. با توجه به دسته‌بندی کاربر و وزن رفتاری او در هر گروه، می‌توان براساس معماری ارائه شده، درخواست او را مورد ارزیابی قرار داده و مجاز و یا غیر مجاز بودن درخواست او را مشخص نمود.

در تحلیل رفتار کاربر باید به مواردی که باعث می‌شود یک دسترسی، به اشتباه مجاز و یا غیر مجاز اعلام شود، توجه کرد. وزن‌دهی کلمات تا حدی باعث می‌شود درخواست‌های هنجاری که دارای کلمات کلیدی ممنوعه با وزن خفیف می‌باشند، وزن رفتاری کاربر را در گروه ناهنجار بالا نبرده و حتی رفتار آرام او، وزن را کاهش می‌دهد. باید یک حد آستانه برای مجاز بودن دسترسی، در بانک قوانین تعریف کرد که براساس آن، نوع مجوز دستیابی بر اساس اطلاعات تحلیل رفتار کاربر تعیین می‌گردد.

۹. بررسی نتایج یک آزمایش

برای ارزیابی معماری ارائه شده و بررسی نتایج آن، ابتدا باید داده‌های اولیه مناسبی تهیه گردد. به همین منظور داده‌هایی به سه صورت منتخب، تصادفی و واقعی فراهم، مورد آزمایش و نتایج آن، بررسی و مورد تحلیل قرار گرفت. در ادامه، این نتایج بیان می‌شوند.

برای بررسی بهتر ابتدا برای پنج کاربر فرضی داده‌هایی تولید شده و بر اساس آن، رفتار کاربر محاسبه گردید. جدول (۶) این داده‌های فرضی را نمایش داده است. کاربر U۱ یک کاربر فرضی است که در ۱۵ درخواست پیاپی خود از هیچ کلمه ممنوعه‌ای استفاده نکرده است. لذا وزن کلمه‌های کلیدی مورد استفاده، صفر منظور می‌شود.

کاربر فرضی U۲ کاربری است که در همه ۱۵ درخواستش از حداقل ۳ کلمه کلیدی ممنوعه با وزن ۱۰۰ استفاده نموده است. کاربر U۳ کاربری است که داده‌های آن به صورت دستی و

جدول ۶. داده‌های ایجاد شده برای ۵ کاربر فرضی

۰ ۰ ۰	۰ ۰ ۰	۰ ۰ ۰	۰ ۰ ۰	۰ ۰ ۰	۰ ۰ ۰	۰ ۰ ۰	۰ ۰ ۰	۰ ۰ ۰	۰ ۰ ۰	۰ ۰ ۰	۰ ۰ ۰	۰ ۰ ۰	۰ ۰ ۰	۰ ۰ ۰	U۱
۱۰۰ ۱۰۰ ۱۰۰	۱۰۰ ۱۰۰ ۱۰۰	۱۰۰ ۱۰۰ ۱۰۰	۱۰۰ ۱۰۰ ۱۰۰	۱۰۰ ۱۰۰ ۱۰۰	۱۰۰ ۱۰۰ ۱۰۰	۱۰۰ ۱۰۰ ۱۰۰	۱۰۰ ۱۰۰ ۱۰۰	۱۰۰ ۱۰۰ ۱۰۰	۱۰۰ ۱۰۰ ۱۰۰	۱۰۰ ۱۰۰ ۱۰۰	۱۰۰ ۱۰۰ ۱۰۰	۱۰۰ ۱۰۰ ۱۰۰	۱۰۰ ۱۰۰ ۱۰۰	۱۰۰ ۱۰۰ ۱۰۰	U۲
۸۰ ۹۰ ۷۰	۹۰ ۹۰ ۵۰	۹۰ ۸۰ ۰	۱۰۰ ۰ ۰	۱۰۰ ۰ ۰	۴۰ ۷۰ ۲۰	۶۰ ۳۰ ۱۰	۵۰ ۲۰ ۰	۳۰ ۰ ۰	۳۰ ۰ ۰	۹۰ ۹۰ ۹۰	۶۰ ۷۰ ۸۰	۹۰ ۶۰ ۷۰	۰ ۰ ۰	۰ ۰ ۰	U۳
۵۰ ۵۷ ۱۶	۴۶ ۲ ۷۴	۳۹ ۳۵ ۲۸	۹۰ ۸۵ ۵۹	۳۱ ۸۰ ۳۱	۱۲ ۶۱ ۷۰	۷۶ ۷۴ ۷۶	۳۳ ۴۵ ۸۳	۵۵ ۳۳ ۵	۲۵ ۶ ۷۳	۱۷ ۶۵ ۷۵	۲۰ ۵۳ ۶۰	۶۸ ۸۶ ۷۵	۱۶ ۵ ۵۶	۱۲ ۵۹ ۶۰	U۴
۴۵ ۷۴ ۶۲	۹۲ ۶۲ ۹۴	۳۷ ۷۲ ۶۱	۴۹ ۸۷ ۱۶	۵۴ ۲۳ ۵۶	۸۳ ۷۹ ۹۸	۸۵ ۲۸ ۱۴	۸۵ ۹۷ ۷۹	۱۷ ۱۴ ۱۹	۵۱ ۵۴ ۹۵	۶۴ ۲۵ ۵۲	۵۵ ۳ ۸۸	۱۲ ۵۶ ۷۷	۹۰ ۳۶ ۹۱	۹۲ ۴۴ ۷۱	U۵

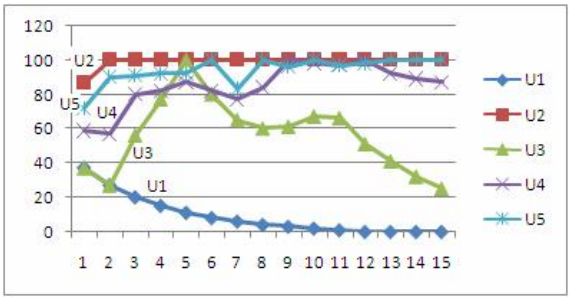
انتخابی تولید شده و کاربرهای U۴ و U۵ کاربرهایی هستند که به هر سه درخواست ایشان یک مقدار تصادفی نسبت داده شده است. محاسبه وزن رفتاری این ۵ کاربر بر اساس فرمول (۲) در جدول (۷) آورده شده است.

جدول ۷. رفتار محاسبه شده برای ۵ کاربر فرضی

۰	۰	۰	۰	۱	۲	۳	۴	۶	۸	۱۱	۱۵	۲۰	۲۷	۳۷	U۱
۱۰۰	۱۰۰	۱۰۰	۱۰۰	۱۰۰	۱۰۰	۱۰۰	۱۰۰	۱۰۰	۱۰۰	۱۰۰	۱۰۰	۱۰۰	۱۰۰	۸۷	U۲
۱۰۰	۹۶	۷۷	۶۶	۶۶	۶۷	۶۱	۶۰	۶۵	۸۰	۱۰۰	۷۷	۵۶	۲۷	۳۷	U۳
۸۷	۸۹	۹۲	۱۰۰	۹۷	۹۸	۱۰۰	۸۴	۷۷	۸۲	۸۷	۸۲	۸۰	۵۷	۵۹	U۴
۱۰۰	۱۰۰	۱۰۰	۹۸	۹۷	۱۰۰	۹۶	۱۰۰	۸۳	۱۰۰	۹۲	۹۲	۹۱	۹۰	۷۲	U۵

نمودار (۱)، نمودار این داده‌ها را بر اساس وزن رفتاری کاربر نسبت به زمان، مشخص کرده است. منحنی‌های پایینی و بالایی نمودار (۱) در واقع حد پایینی و حد بالایی داده‌ها را مشخص ساخته است. منحنی U۱ مربوط به کاربری است که هیچ کلمه ممنوعه‌ای به کار نبرده است و لذا منحنی کم‌کم به سمت صفر میل کرده و منطبق بر خط صفر ادامه یافته است. یعنی U۱ یک کاربر با رفتار کاملاً هنجار می‌باشد.

منحنی U۲ مربوط به کاربری است که بر خلاف کاربر اول، رفتاری کاملا ناهنجار داشته و به سرعت بر خط ۱۰۰ منطبق شده است. نمودار U۳ کاربری است که پس از ارتباط اولیه، رفتارهای ناهنجاری داشته و پس از چند دسترسی رفتارش اصلاح شده و به صورت متعادل درآمده است. منحنی‌های U۴ و U۵ مربوط به کاربران با رفتار تصادفی است. البته بهتر است گفته شود با رفتار ناهنجار تصادفی، زیرا برای هر یک، حتما سه مقدار تصادفی فرض کرده که عددی بین ۰ تا ۱۰۰ است. بنابراین کاربرانی که رفتار ناهنجار خود را به هر شکل ادامه دهند، شناسایی شده و نمودار آن‌ها بر خط ۱۰۰ منطبق می‌شود.



نمودار ۱. نمودار رفتاری ۵ کاربر فرضی

جدول ۸. داده‌های ۵ کاربر واقعی

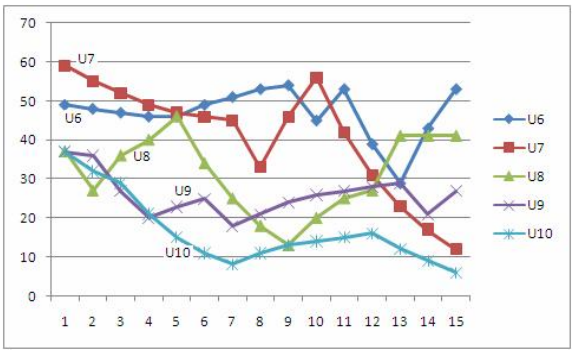
۶۰ ۷۰ ۰	۶۰ ۷۰ ۰	۰ ۰ ۰	۰ ۰ ۰	۲۰ ۳۰ ۷۰	۳۰ ۰ ۰	۲۰ ۷۰ ۰	۲۰ ۷۰ ۰	۲۰ ۷۰ ۰	۲۰ ۷۰ ۰	۷۰ ۰ ۰	۷۰ ۰ ۰	۷۰ ۰ ۰	۷۰ ۰ ۰	۷۰ ۰ ۰	U۶
۰ ۰ ۰	۰ ۰ ۰	۰ ۰ ۰	۰ ۰ ۰	۰ ۰ ۰	۰ ۰ ۰	۶۰ ۷۰ ۷۰	۰ ۰ ۰	۶۵ ۰ ۰	۶۵ ۰ ۰	۶۵ ۰ ۰	۶۵ ۰ ۰	۶۵ ۰ ۰	۶۵ ۰ ۰	۶۰ ۷۰ ۰	U۷
۶۵ ۰ ۰	۶۵ ۰ ۰	۶۵ ۶۵ ۰	۵۵ ۰ ۰	۶۵ ۰ ۰	۶۵ ۰ ۰	۰ ۰ ۰	۰ ۰ ۰	۰ ۰ ۰	۰ ۰ ۰	۱۰۰ ۰ ۰	۸۰ ۰ ۰	۱۰۰ ۰ ۰	۰ ۰ ۰	۰ ۰ ۰	U۸
۷۰ ۰ ۰	۰ ۰ ۰	۵۰ ۰ ۰	۵۰ ۰ ۰	۵۰ ۰ ۰	۵۰ ۰ ۰	۵۰ ۰ ۰	۵۰ ۰ ۰	۰ ۰ ۰	۵۰ ۰ ۰	۵۰ ۰ ۰	۰ ۰ ۰	۰ ۰ ۰	۵۰ ۰ ۰	۰ ۰ ۰	U۹
۰ ۰ ۰	۰ ۰ ۰	۰ ۰ ۰	۰ ۰ ۳۰	۰ ۰ ۳۰	۰ ۰ ۳۰	۰ ۰ ۳۰	۰ ۰ ۰	۰ ۰ ۰	۰ ۰ ۰	۰ ۰ ۰	۰ ۰ ۰	۰ ۰ ۳۰	۰ ۰ ۳۰	۰ ۰ ۰	U۱۰

برای بررسی بیشتر، اطلاعات دسترسی به صفحات وب مربوط به ۵ کاربر از فایل گزارش‌های سایت AOL انتخاب شده و داده‌های آن‌ها در جدول (۸) آورده شده است. محاسبه رفتار ایشان در جدول (۹) و نمودار رفتاری آن‌ها در نمودار (۲) آمده است.

جدول ۹. وزن رفتار محاسبه شده برای ۵ کاربر واقعی

۵۳	۴۳	۲۹	۳۹	۵۳	۴۵	۵۴	۵۳	۵۱	۴۹	۴۶	۴۶	۴۷	۴۸	۴۹	U۶
۱۲	۱۷	۲۳	۳۱	۴۲	۵۶	۴۶	۳۳	۴۵	۴۶	۴۷	۴۹	۵۲	۵۵	۵۹	U۷
۴۱	۴۱	۴۱	۲۷	۲۵	۲۰	۱۳	۱۸	۲۵	۳۴	۴۶	۴۰	۳۶	۲۷	۳۷	U۸
۲۷	۲۱	۲۹	۲۸	۲۷	۲۶	۲۴	۲۱	۱۸	۲۵	۲۳	۲۰	۲۷	۳۶	۳۷	U۹
۶	۹	۱۲	۱۶	۱۵	۱۴	۱۳	۱۱	۸	۱۱	۱۵	۲۱	۲۹	۳۲	۳۷	U۱۰

بر طبق نمودار رفتار کاربران، کاربر U۶ از ابتدا رفتار ناهنجار داشته و سعی در ادامه آن نیز دارد. اگر در زمان‌هایی نیز به سمت هنجار تمایل پیدا کرده، سریعاً به حالت ناهنجار باز گشته است. کاربر U۷ با رفتار ناهنجار شروع کرده و پس از مدتی دارای رفتار هنجار شده است. کاربر U۸ رفتار نامشخصی داشته و به صورت شبه‌سینوسی بین هنجار و ناهنجار نوسان می‌کند. کاربر U۹ رفتاری متعادل داشته و کاربر U۱۰ کاربری آرام بوده است. بنابراین می‌توان گفت چنین کاربرانی همزمان دو یا چند موضوع را پیگیری می‌کنند و لذا رفتار ایشان نامشخص است.



نمودار ۲. نمودار رفتاری ۵ کاربر واقعی

در این آزمایش از داده‌های کاربران با رفتارهای افراطی و یا بسیار آرام صرف نظر شده است. در این مقاله مجال بررسی داده‌های بیشتر برای تحلیل مسایلی از قبیل کاربران چند رفتاری نمی‌باشد. کاربران چندرفتاری می‌توانند به‌عنوان یک موضوع جداگانه در تحقیقات آتی مورد بررسی قرار گیرد.

۱۰. چالش‌های فرارو

اگر چه پالایش وب با استفاده از تحلیل رفتار کاربران می‌تواند کارایی فیلترها را بهبود بخشد، ولی با چالش‌هایی روبرو می‌باشد که در ادامه تشریح می‌شود.

۱۰-۱. سرعت دستیابی

با توجه به اینکه سیستم فیلترینگ در مسیر اصلی ارتباط با اینترنت قرار داده می‌شود و درخواست‌های همه کاربران باید با مجوز این سیستم پاسخ داده شود، سرعت دستیابی به صفحه مورد نظر از اهمیت ویژه‌ای برخوردار خواهد بود. پردازش‌های مربوط به کلاس‌بندی کاربران، استخراج الگوی رفتار و تحلیل آن و تعیین مجوز دستیابی کاربر، زمان زیادی صرف نموده و باعث تاخیر چشمگیر در پاسخ به کاربران می‌شود.

اگر در هنگام پیاده‌سازی به مسایل فوق توجه نگردد، ممکن است سیستم شبکه، ناکارآمد شود. طبق معماری ارائه شده، مجوز عبور درخواستی کاربر، بر اساس اطلاعات بانک رفتار وی صادر می‌گردد. با توجه به اینکه هنگام درخواست، استخراج الگو و کلاس‌بندی کاربران صورت نمی‌پذیرد، پس این تاخیرها شامل زمان پاسخ نمی‌شود. اما تطبیق کاربر با نشست مربوطه و دسترسی به بانک رفتار همچنان جزء مسایلی خواهد بود که باید در پیاده‌سازی، آن را به حداقل رساند.

۱۰-۲. خطا در گذر از فیلتر

از دیگر چالش‌های مطرح، اشتباه در مسدود ساختن یک صفحه و یا صدور مجوز عبور می‌باشد. در روش ارائه شده، رفتار کاربر با مجموعه‌ای از عوامل در حین درخواست و سابقه رفتار و نیز وضعیت او در درخواست جدید سنجیده می‌شود. ممکن است این نتیجه بنا به دلایلی از جمله به‌روز نبودن بانک کلمات کلیدی ممنوعه و یا عدم دقت در تعیین ضرایب مربوط به فرمول رفتار کاربر، با واقعیت تفاوت داشته باشد.

۱۰-۳. وزن‌دهی کلمات کلیدی ممنوعه

در وزن‌دهی کلمات ممنوعه، مشکل اصلی، چند ماهیتی بودن کلمات می‌باشد. یعنی یک کلمه ممکن است به‌عنوان یک واژه ممنوعه به‌کار برده شده ولی به عنوان یک کلمه عادی استفاده شود. وزن‌دهی کلمات می‌تواند تا حدی ماهیت کلمات را مشخص نماید. اگر کلمات را به‌صورت صحیحی وزن‌دهی نماییم و با توجه به اینکه نوعی میانگین‌یابی در تعیین وزن‌ها وجود دارد، تا حدی باعث تاثیر ماهیت کلمات در وزن‌های به‌دست آمده می‌شود.

به هر حال، راه حل فوق یک راه حل نسبی بوده و این موضوع می‌تواند به‌عنوان یک موضوع مستقل در تحقیقات آتی پیگیری شود.

۱۱. نتیجه‌گیری

هدف در این مقاله استفاده از تحلیل رفتار کاربران برای بهبود مسدودسازی درخواست‌های ممنوعه است. به همین منظور با استفاده از اطلاعات فایل گزارش دسترسی کاربران و داده‌کاوی آن، الگوی رفتار ایشان استخراج شده و در یک بانک اطلاعات

ذخیره می‌گردد. برای کمّی کردن رفتار کاربر، قوانینی ارائه شد که در آن، عواملی مانند وزن کلمات کلیدی مورد استفاده و همچنین سابقه رفتار کاربر تعیین‌کننده هستند. بر اساس نتایج بدست آمده از این محاسبات، وزن رفتاری کاربر تعیین شده و در بانک اطلاعات مربوطه ذخیره می‌گردد. در هر دسترسی کاربر، ضمن استفاده از این اطلاعات به عنوان سابقه رفتار وی، بر اساس دسترسی جدید، اطلاعات فوق به‌روز شده و ذخیره می‌گردد. با مراجعه به این بانک و بررسی واژه‌های به کار رفته توسط کاربر، تقاضای او تحلیل و ارزیابی گشته و مجوز عبور صادر شده و یا صفحه درخواست شده مسدود می‌گردد.

دوری از مسدود کردن افراطی صفحات وب، مزیت اصلی روش ارائه شده، نسبت به روش معمول مورد استفاده در فیلترها است.

بررسی چگونگی جلوگیری از اشتباهات ناشی از ماهیت کلمات کلیدی و اشتباه در گذر از فیلتر و ارائه راه‌کار مناسب، می‌تواند جزء کارهایی باشد که در آینده به بهبود پالایش صفحات منجر شود. اصلاح وزن‌دهی بر اساس اطلاعات جانبی و به‌روز کردن خودکار آن می‌تواند یک رویکرد باشد.

ارائه یک معماری که در آن با استفاده از اطلاعات دیگری مانند کلمات غیر ممنوعه که به نحوی در وزن‌دهی اثر منفی داشته و با استفاده از آن بتوان رفتارهای حرفه‌ای را، مانند آنچه در مورد جستجوی یک پزشک گفته شد، تشخیص داده و باعث بهبود درک رفتار شود، می‌تواند جزء کارهای آتی باشد.

اصلاح هوشمندانه وزن‌دهی کلمات کلیدی بر اساس رفتار کاربر می‌تواند به‌عنوان محور تحقیقی جدیدی مطرح و پیگیری گردد. همان‌طور که قبلا بیان شد، در ابتدا باید در جدول وزن‌دهی کلمات، به هر کلمه کلیدی وزنی متناسب کاربرد آن در جستجوهای ممنوعه اختصاص یابد که این‌کار به صورت دستی و بر اساس بررسی و تحلیل مدیر سیستم فیلتر صورت می‌پذیرد و می‌توان ضرایب کلمات کلیدی را بر اساس رفتار کاربر و رهگیری او اصلاح کرده و کم و یا زیاد نمود.

برای بالا بردن دقت تعیین الگوی رفتار کاربر می‌توان پارامتر زمان را نیز منظور نمود. تناوب درخواست‌ها و یا فاصله زمانی آن‌ها می‌تواند ملاکی بر اصرار و یا کم‌توجهی کاربر بر ادامه رفتار خود باشد. با استفاده از فیلد زمان درخواست کاربر در فایل گزارش، می‌توان سنجه‌های جدیدی برای اصلاح تعیین الگوی رفتار طراحی کرده و استفاده نمود.

برای بالا بردن کارایی بیشتر فیلترها، می‌توان به جای استفاده از مجوز عبور و یا عدم عبور، بر اساس درخواست کاربر و سابقه رفتاری وی مجوزهای خاصی را صادر نمود. یعنی براساس طیف اخلاقی کابران، مجوزهای معنی‌داری را صادر کرد. به‌عنوان مثال کاربرانی با سابقه رفتاری ناهنجار با نمره ۸۰ الی ۹۰ اجازه دیدن متن را داشته ولی مجوز دیدن فایل‌های چند رسانه‌ای را نداشته باشند. این امر کارایی فیلترها را بیشتر کرده و هدفمندتر می‌گردد.

۱۳. مراجع

[1] WebSpy Ltd, ."Internet Use Statistic.", Internet: http://www.webspy.com/resources/whitepapers/2008%20WebSpy%20Ltd%20%20Internet%20Use%20Statistics.pdf, 2008.

[2] Connect World, "Employee Internet Usage."; Internet: http://www.connectworld.net/ccc/employee-internet-usage. html, It is a snapshot of the page as it appeared on 14 Dec 2010.

[3] Covenant Eyes, "Internet Pornography Statistics."; Internet: http://www.covenanteyes.com/support/pdfs/Covenant%20Eyes%20Pornography%20Statistics.pdf, 2006

[4] Yaltaghian, Behnak "Improving the ranking of Search Engine Output: A Network Analysis Approach."; In IBM Centre for Advanced Studies Conferenc, University of Toronto, Toronto, Ontario, M5S 2E4, 2002.

[5] Bathumalai, Ganesan "Self Adapting Websites: Mining User Access Logs."; M. Sc. Thesis, The Robert Gordon University, 2008.

[6] Kosala, Raymond; Blockeel, Hendrik "Web Mining Research: A Survey."; In ACM SIGKDD Explorations, 2000, 2(1), 1-15.

[7] Mobasher, Bamshad "Web usage mining."; In Bing Liu, Web Data Mining: Exploring Hyperlinks, Content and Usage Data, Berlin-Heidelberg, Springer-Verlag 2007, 449-483.

[8] Hofgesang, I. Peter "Web Usage Mining Structuring Semantically Enriched Clickstream Data."; M. Sc. Thesis, Vrije University Amsterdam, The Netherlands, 2004.

[9] Mihara, Koichiro; Terabe, Masahiro; Hashimoto, Kazuo "A Novel Web Usage Mining Method."; In Proceedings of the Fourth International Conference on Web Information Systems and Technologies 2008.

[10] Xu, Guandong "Web Mining Techniques for Recommendation and Personalization."; Ph. D. Thesis, Victoria University, Australia, 2008.

[11] Liu, Ning-Han; Lai, Szu-Wei; Chen, Chien-Yi; Hsieh, Shu-Ju "Adaptive Music Recommendation Based on User Behavior in Time Slot."; In International Journal of Computer Science and Network Security 2009, 9(2), 219-227, Available on: http://paper.ijcsns.org/07_book/ 200902/ 20090229.pdf