تعامل انسان و اطلاعات، جلد۴ شماره ۲
http://hii.khu.ac.ir
مقاله اصیل پژوهشی

طراحی سیستمی به‌منظور تحلیل روند مراجعه‌ی کاربران اینترنتی به وب‌سایت‌ها در ایران با استفاده از الگوریتم‌های داده‌کاوی و متن‌کاوی

بابک سهرابی: استاد مدیریت فناوری اطلاعات، دانشکده مدیریت، دانشگاه تهران، تهران، ایران (نویسنده مسئول). bsohrabi@ut.ac.ir
ایمان رئیسی وانانی: استادیار مدیریت صنعتی، دانشکده مدیریت و حسابداری، دانشگاه علامه طباطبائی، تهران، ایران.
محمدرضا خرمی: فارغ التحصیل کارشناسی ارشد مدیریت فناوری اطلاعات، دانشکده مدیریت، دانشگاه تهران، تهران، ایران.

دریافت: ۱۳۹۶/۶/۲۸
پذیرش: ۱۳۹۶/۸/۲۸

چکیده

زمینه و هدف: با توجه به ورود وب‌گردی به سبک زندگی طیف وسیعی از افراد جامعه و احساس نیاز به سیاست‌گذاری‌های فرهنگی و اجتماعی دقیق‌تر در این حوزه، محققین پژوهش حاضر بر آن شدند تا با تحلیل رفتار افراد جامعه در مراجعه به وب‌سایت‌های مختلف، سیاست‌گذاران را در تصمیم گیری کارآمدتر یاری نمایند.

روش پژوهش: در این تحقیق از گام های روش علم طراحی استفاده شده است. جامعه‌ی داده این تحقیق، کل کاربران بازدیدکننده از وب‌سایت‌ها در ایران هستند که در دسترس بوده اند. برای انجام این تحقیق نیاز به اطلاعات وب‌گردی افرادی از صنوف مختلف جامعه وجود داشت که بدین منظور با طراحی و انتشار افزونه‌هایی قابل نصب بر روی مرورگرهای مختلف، داده‌های مورد نیاز جمع‌آوری گردید.

یافته‌ها: با استفاده از الگوریتم‌های متن‌کاوی، صفحات رجوع‌شده از منظر محتوا از هم تفکیک شده و سپس به کمک الگوریتم‌های داده‌کاوی، مراجعات کاربران و همچنین صفحات اینترنتی دسته‌بندی شده و هر دسته تفسیر شده است. در نهایت با توجه به داده‌های جمع‌آوری‌شده و دسته‌بندی‌های انجام شده، گزارشات متنوع و کارآمدی به عنوان نمونه و با توجه به نیاز جامعه تصمیم گیرنده، آماده و ارائه گردیده است.

نتیجه‌گیری: در نهایت یک سیستم جامع به منظور تحلیل روند وب گردی کاربران اینترنتی طراحی گردید که از جمع آوری اطلاعات تا آماده سازی گزارشات نوآورانه را در بردارد که می تواند به عنوان یک نمونه ای کارآمد جهت تحلیل، طراحی و پیاده سازی سیستم‌های تحلیلی تحت وب مورد استفاده قرار گیرد.

کلیدواژه‌ها: دسته‌بندی صفحات اینترنتی، خوشه‌بندی مراجعات، وب‌گردی، تحلیل روند

مقدمه

امروزه تعداد کاربران اینترنت از مرز سه میلیارد گذشته‌اند و این تعداد به صورت اعجاب‌آوری در حال افزایش است[1]. این آمار در کنار آن‌چه در اطراف خود می‌بینیم به خوبی نشان می‌دهد که اینترنت وارد زندگی مردم شده و تعامل با آن جزئی از سبک زندگی امروز جهان ماست؛ پس نیاز است تا به این مقوله بیشتر بها داده شود. از طرفی حکومت به‌منظور سیاست‌گذاری در فضای فرهنگی–اجتماعی جامعه، نیاز به رصد این فضا دارد. امروزه قطعا یکی از مولفه‌های غیرقابل انکار این فضا، فضای مجازی بوده و از مهمترین ابعاد آن شبکه‌های اجتماعی و وب‌سایت‌ها هستند. از طرفی خوشبختانه در حال حاضر در کشور ما پژوهش‌های متعددی روی شبکه‌های اجتماعی و رفتار افراد جامعه در آن انجام شده و در حال انجام است، ولی در مورد پژوهش روی رفتار افراد در تعامل با وب‌سایت‌های اینترنتی کمبودهایی دیده می‌شود.

همچنین امروزه برخی از دست‌اندرکاران جوامع کوچک و بزرگ دیگر نیز علاقه‌مندند تا تعاملات اینترنتی افراد جامعه‌ی خود را رصد نمایند. به‌عنوان مثال امروزه در کشور ما این دیدگاه در بین مردم رایج شده که کارمندان (به خصوص کارمندان بخش دولتی) در ساعات کاری به‌جای انجام کارهای محوله، گاها به اتلاف وقت و تفریح در فضای مجازی –که یکی از موارد آن وب‌گردی می‌باشد– می‌پردازند. به منظور بررسی این موضوع، نیاز است تا وب‌گردی آن‌ها رصد شده و بررسی شود که بیشتر در چه فضاهایی گردش می‌کرده‌اند. حال می‌توان با طراحی سیستمی جامع، این

[1] آمار لحظه‌ای تعداد کاربران اینترنتی را می‌توانید در www.internetlivestats.com مشاهده کنید.

امکان را برای دولت و سازمان‌های متنوع دولتی و خصوصی فراهم کرد تا با پیاده‌سازی این سیستم به سادگی بتوانند روند وب‌گردی نیروی انسانی خود را در ساعات کاری تحلیل کرده و تصمیمات لازم را اتخاذ نمایند.

در کنار مسائل بالا، یکی از مسائلی که در کشور ما و به‌خصوص در سالیان اخیر هرازگاهی مطرح شده بحث فیلترینگ، و شدت و شکل آن است. نظرات گوناگون در این باب، هر کدام استدلال‌هایی دارند ولیکن به‌دلیل عدم وجود داده‌های واقعی هیچ‌کدام به طور قاطع گسترش نمی‌یابد. در این تحقیق می‌توان در کنار پاسخ به مسائل بالا، داده‌هایی واقعی از استفاده از فیلترشکن و *VPN* جمع‌آوری کرده تا با تحلیل آن‌ها، اطلاعاتی در باب هدف و شکل استفاده‌ی کاربران مختلف از فیلترشکن و *VPN* استخراج نمود. این اطلاعات می‌تواند در اختیار نهادهای مربوطه قرار گرفته و آن‌ها را در سیاست‌گذاری در مقوله‌ی فیلترینگ یاری رساند.

این مسائل در گستره‌ی جهانی مطرح هستند، لکن به دلیل محدودیت‌های زمانی و مالی، به دلیل این‌که این تحقیق جنبه‌ی پژوهشی دارد و همچنین به‌دلیل اقتضائات فرهنگی خاص کشورمان، در این پژوهش قلمرو مسئله به کاربران اینترنتی در ایران محدود شده است.

مبانی نظری و پیشینه تحقیق

مروری بر مبانی نظری تحقیق: کابران اینترنت به طرق مختلفی از اینترنت استفاده می‌کنند. یکی از متداول‌ترین روش‌های تعامل با اینترنت، مراجعه به وب‌سایت‌های اینترنتی از طریق مرورگرهای وب[1] توسعه‌داده‌شده بر روی رایانه‌های شخصی، لپ‌تاپ‌ها، تبلت‌ها و تلفن‌های همراه هوشمند می‌باشد. به علت تنوع بسیار زیاد وب‌سایت‌های اینترنتی، این مراجعه با اهداف کاملا متفاوتی ممکن است انجام شود از قبیل سرگرمی، بازی، انجام فرآیندهای اداری، خبرخوانی، تعامل و ارتباط، مطالعات علمی، آموزش، خرید، دانلود و موارد مشابه.

در این مسیر، طراحی سیستم و تحلیل های مدیریتی در واقع فرآیندی است که با تعریف کردن معماری سیستم، ماژول‌ها، مولفه‌ها و ارتباطات آن‌ها، یک سیستم را به منظور رفع نیازمندی‌های خاصی معرفی می‌کند. این تحقیق معماری، ماژول‌ها و مولفه‌ها و ارتباطات سیستمی را ارائه کرده است که با پیاده‌سازی آن می‌توان روند مراجعه‌ی کاربران اینترنتی به وب‌سایت‌ها را کشف و تحلیل نمود. به منظور تحلیل داده‌هایی که از طریق سیستم ها، اینترنت و فایل های متعدد به دست می آیند، نیاز است که از الگوریتم ها و روش های پیشرفته تحلیلی استفاده گردد که در ادامه، به دو گروه مهم از آنها اشاره می گردد.

الگوریتم‌های داده‌کاوی[2] و متن‌کاوی[3]: داده‌کاوی به‌عنوان یک گام در فرآیند کشف دانش شناخته می‌شود؛ گامی اساسی که در آن الگوهای نهان آشکار و ارزیابی می‌شوند. به هر حال در صنعت، در رسانه‌ها و در محیط‌های تحقیقاتی و پژوهشی اغلب واژه‌ی داده‌کاوی به کل فرآیند کشف دانش اشاره می‌کند. (Han, et al., 2011).

در حقیقت، داده‌کاوی به مفهوم استخراج اطلاعات نهان، الگوها و روابط در حجم انبوهی از داده‌هاست. تکنیک‌های داده‌کاوی پایگاه داده‌ها و مجموعه‌های حجیمی از داده‌ها را در پی کشف و استخراج دانش مورد کندوکاوهای ماشینی قرار می‌دهد. (Fayyad & Piatetsky-Shapiro, 1996) این الگوریتم‌ها بر اساس اهداف و کارکردشان در چند دسته قرار می‌گیرد، از جمله قواعد انجمنی، ترتیب، پیش‌بینی، طبقه‌بندی، خوشه‌بندی و مصورسازی. (Larose, 2014).

با گذر زمان و به خصوص با فراگیر شدن اینترنت، نیاز به کاوش و استخراج دانش از داده‌های نیمه‌ساخت‌یافته و غیرساخت‌یافته (از جمله متون) احساس شد. بدین منظور لازم بود تا ابتدا این داده‌ها را ساختارمند نمود و در ادامه از تکنیک‌های داده‌کاوی به منظور کشف دانش استفاده کرد. متن‌کاوی همین مراحل را به منظور کشف دانش از متون انجام می‌دهد و به لحاظ مفهوم روش های مورد استفاده، با داده کاوی اشتراک و غرابت دارد (Niknafs & Niknam, 2016).

در این تحقیق ابتدا از الگوریتم‌های متن‌کاوی به‌منظور تفکیک محتوایی صفحات اینترنتی مورد رجوع جامعه‌ی آماری استفاده شده و در ادامه نیز از الگوریتم‌های داده‌کاوی ازجمله خوشه‌بندی به‌منظور دسته‌بندی جامعه‌ی آماری بر اساس روند مراجعه به وب‌سایت‌ها استفاده شده است.

الگوریتم مدل فضای برداری[4]: این الگوریتم را می‌توان سبک‌ترین و ساده‌ترین الگوریتم دسته‌بندی متون نامید. این روش به طور کلی بر این اصل استوار است که برای هر دسته از متون یک بردار یکه می‌سازد طوری‌که محورهای این بردار

[1] Web browser

[2] Data mining

[3] Text mining

[4] Vector Space Model

تکه[1]ها (کلمات) متن هستند و طول بردار در راستای هر بعد نیز نسبتی از تعداد تکرار تکه (کلمه) در متن است. سپس برای متن جستار نیز یک بردار می‌سازد و در انتها نزدیکترین بردار به بردار جستار را پیدا کرده (مثلا از طریق پیدا کردن بیشترین کسینوس زاویه) و جستار را به آن دسته نسبت می‌دهد. (*Salton, et al., 1975*). در ادامه برخی از الگوریتم های مدل فضای برداری برای تحلیل متون شرح داده شده و از آنها استفاده خواهد شد.

الگوریتم ***K-means***: در سال ۱۹۷۵ هارتیگان برای اولین بار الگوریتم *K-means* را ارائه داد و در سال ۱۹۷۹ به کمک ونگ تغییراتی در آن ایجاد کرد و اکنون متداول‌ترین ابزار خوشه‌بندی استفاده‌شده در کاربردهای صنعتی و علمی است. در این روش، خوشه‌ها با مراکزشان که معمولا میانگین نقاط درون یک خوشه است، بیان می‌شوند. در این روش فاصله‌ی هر نقطه تا مرکز آن خوشه، به‌عنوان تابع هدف در نظر گرفته می‌شود. فاصله می‌تواند تعاریف و گستره‌ای را شامل شود. هر نقطه به خوشه‌ای تعلق دارد که به مرکز جرم آن نزدیک‌تر است. در این روش تعداد خوشه‌ها (k) باید مشخص باشد. کلیت الگوریتم بدین صورت است:

- انتخاب k نقطه به‌عنوان مراکز اولیه‌ی خوشه‌ها
- تخصیص هر نقطه به خوشه‌ای که به مرکز آن خوشه نزدیک‌تر است
- محاسبه‌ی مجدد مراکز خوشه‌ها تا جایی‌که تکرار مراحل یادشده، تغییری در خوشه‌ها و مراکز آن‌ها ایجاد نکند. (*Abtahi, et al., 2017*)

الگوریتم ***K-modes***: این الگوریتم در واقع همان الگوریتم *k-means* است که برای داده‌های تماما دسته‌ای[2] تعریف شده و استفاده می‌شود. (*Huang, 1998*) قابل ذکر است این الگوریتم برای خوشه‌بندی داده‌های با حجم بالا، بهترین گزینه است.

الگوریتم ***K-prototype***: معتبرترین و پرکاربردترین الگوریتم برای خوشه‌بندی داده‌هایی که هم فیلد عددی[3] دارند و هم فیلد دسته‌ای، الگوریتم *k-prototype* است. (*Huang, 1997*) از معایب این الگوریتم این است که اولا تعداد خوشه‌ها را به عنوان ورودی خوشه‌بندی می‌گیرد و دوما معیار مشخص و تعریف‌شده‌ای برای امتیازدهی به خوشه‌بندی انجام شده ارائه نکرده است.

الگوریتم ***PCL-OC***[4]: این الگوریتم نیز برای خوشه‌بندی داده‌هایی که هم فیلد عددی دارند و هم فیلد دسته‌ای، استفاده می‌شود. این الگوریتم علاوه بر داده‌های خوشه‌بندی، یک عدد به عنوان حداکثر تعداد خوشه به عنوان ورودی می‌گیرد و عملیات خوشه‌بندی را انجام می‌دهد. البته قابل ذکر است که این الگوریتم برای تشخیص تعداد خوشه، تنها کاری که می‌کند این است که خوشه‌بندی را با حداکثر تعداد خوشه‌ی مشخص‌شده انجام داده و سپس اگر تعداد اعضای خوشه‌ای از یک حداقلی کمتر باشد، آن خوشه را با نزدیک‌ترین خوشه ادغام خواهد کرد. با این حال در مواردی که تعداد رکورد داده‌ی بسیار زیاد و تعداد پارامتر کم باشد ادعا شده است که این الگوریتم نسبت به الگوریتم *k-prototype* دقت و سرعت بالاتری دارد. (*Cheung & jia, 2013*).

مروری بر پیشینه‌ی تحقیق

در سال‌های اخیر پژوهش‌های زیادی در باب تفکیک صفحات اینترنتی و هم‌چنین تحلیل روند وب‌گردی انجام شده است که هر یک ویژگی‌های خاص خود را دارند.

از دو دهه پیش تفکیک محتوایی صفحات اینترنتی در میان پژوهشگران حوزه‌ی داده‌کاوی و متن‌کاوی مطرح بوده است. در ادامه به چند مورد از اولین مقالات منتشرشده در این حوزه اشاره شده است:

- ملادنیک[5] در سال ۱۹۹۸ مقاله‌ای تحت عنوان "تبدیل یاهو به یک طبقه‌بندی‌کننده‌ی خودکار صفحات وب[6]" منتشر کرده است. وی در این مقاله با استفاده از مدل فضای برداری و البته با این تفاوت که به جای تکه‌تکه کردن عبارات به صورت معمول[7]، تکه‌های چندتایی[8] تشکیل داده است، موفق شده با کمک ساختار سلسله‌مراتبی مسئله را به مسائل کوچکتری شکسته –به طوری که هر زیر مسئله میزان وابستگی متن به یک مفهوم را مشخص می‌کند– و این گونه احتمال این‌که متن در هر دسته‌ی محتوایی قرار می‌گیرد را پیدا کرده و این‌چنین دسته‌بندی مورد نظرش را انجام داده است (Mladenic, 1998).

---

[1] Token
[2] Categorical
[3] Numeric
[4] Penalized competitive learning based on object-cluster similarity metric
[5] Mladenic
[6] Turning yahoo into an automatic web-page classifier
[7] unigram
[8] N-gram

• اتاردی[۱]، گولی[۲] و سباستینی[۳] در سال ۱۹۹۹ مقاله‌ای تحت عنوان "دسته‌بندی خودکار صفحات وب با استفاده از لینک و تحلیل متن[۴]" منتشر کرده‌اند. آن‌ها در این مقاله در کنار تکنیک‌های متن‌کاوی، از لینک صفحات اینترنتی نیز برای دسته‌بندی آن‌ها استفاده می‌کنند. (Attardi, et al., 1999)

• وون[۵] و لی[۶] در سال ۲۰۰۰ مقاله‌ای تحت عنوان "طبقه‌بندی صفحات وب بر اساس رویکرد k–نزدیک‌ترین همسایه[۷]" منتشر کرده‌اند. این دو محقق در این مقاله با استفاده از روش گزینش ویژگی[۸] و هم‌چنین وزن‌دهی کلمات[۹] در مدل فضای برداری سعی کرده‌اند دقت الگوریتم k–نزدیک‌ترین همسایه را در دسته‌بندی محتوایی متون افزایش دهند. (Kwon & Lee, 2000)

• دامیس[۱۰] و چن[۱۱] در سال ۲۰۰۰ مقاله‌ای منتشر کرده‌اند تحت عنوان "طبقه‌بندی سلسله‌مراتبی محتوای وب[۱۲]". این دو در این پژوهش موفق شده‌اند با استفاده از طبقه‌بندی‌کننده‌های ماشین بردار پشتیبان[۱۳]، محتوای وب را در دو سطح به صورت سلسله‌مراتبی طبقه‌بندی کنند. (Dumais & Chen, 2000)

• پنگ[۱۴] و چوی[۱۵] در سال ۲۰۰۲ مقاله‌ای با عنوان "طبقه‌بندی خودکار صفحات وب به روشی پویا و سلسله‌مراتبی[۱۶]" منتشر نموده‌اند. این دو در این تحقیق قابلیت پویا بودن را به تحقیقات قبلی این حوزه، اضافه کرده‌اند. منظور از پویا بودن در این تحقیق، امکان اضافه شدن دسته‌های جدید به دسته‌بندی صورت گرفته است. (Peng & Choi, 2002) با توجه به تنوع روز افزون صفحات وب، امکان ارائه و تاکیدشده در این مقاله، کاربرد بسیاری خواهد داشت.

با توجه به گسترش روزافزونه صفحات وب و کاربرد وسیع این مسئله –دسته‌بندی محتوایی صفحات وب– ، کماکان این مسئله در حوزه‌ی آکادمیک مطرح است و هر ساله چندین مقاله در این باب منتشر می‌شود. در ادامه‌ی این متن، به چند مقاله‌ی دیگر که در سال‌های اخیر منتشر شده، اشاره شده است:

• چن و سیه[۱۷] در سال ۲۰۰۶ مقاله‌ای با عنوان "طبقه‌بندی صفحات وب براساس ماشین بردار پشتیبان و با استفاده از یک طرح ارزیابی وزنی[۱۸]" منتشر کرده‌اند. این دو در این تحقیق با استفاده‌ی هم‌زمان از تحلیل معنایی نهفته[۱۹] و گزینش ویژگی‌های صفحات وب در کنار مدل ماشین بردار پشتیبان موفق به طراحی سیستمی به منظور طبقه‌بندی صفحات وب شدند. (Chen & Hsieh, 2006)

• ازیل[۲۰] در سال ۲۰۱۱ و در مقاله‌ای که با عنوان "یک سیستم طبقه‌بندی صفحات وب بر اساس الگوریتم ژنتیک و با استفاده از برچسب‌ها به عنوان ویژگی[۲۱]" منتشر کرده است، از تگ‌های HTML و همچنین برچسب‌ها به عنوان ویژگی‌هایی برای طبقه‌بندی صفحات وب استفاده کرده است. وی در این مقاله ادعای دقت بالای ۹۵ درصد و بالاتر از الگوریتم‌های طبقه‌بندی نایو بیز[۲۲] و k–نزدیک‌ترین همسایه[۲۳] کرده است. (Özel, 2011)

• سانوجا[۲۴] و گنکارسکی[۲۵] در سال ۲۰۱۴ مقاله‌ای تحت عنوان "چارچوب تقسیم‌بندی صفحات وب[۲۶]" منتشر کرده‌اند. آن‌ها در این مقاله یک رویکرد ترکیبی استفاده کرده‌اند و هر صفحه‌ی وب را از جهت سه ساختار درخت DOM، ساختار محتوا و ساختار منطق مورد توجه قرار داده‌اند. (Sanoja & Gancarski, 2014)

• سیارلی[۲۷]، الیویرا[۱] و سالس[۲] در سال ۲۰۱۴ مقاله‌ای با

---

۱ Attardi
۲ Gulli
۳ Sebastiani
۴ Automatic Web page categorization by link and context analysis
۵ Kwon
۶ Lee
۷ Web page classification based on k-nearest neighbor approach
۸ Feature Selection
۹ Term-Weighting
۱۰ Dumais
۱۱ Chen
۱۲ Hierarchical classification of Web content
۱۳ Support Vector Machine (SVM)
۱۴ Peng
۱۵ Choi
۱۶ Automatic web page classification in a dynamic and hierarchical way
۱۷ Hsieh
۱۸ Web page classification based on a support vector machine using a weighted vote schema
۱۹ Laten semantic analysis
۲۰ Özel
۲۱ A web page classification system based on a genetic algorithm using tagged-terms as features.
۲۲ Naïve Bayes
۲۳ K-Nearest Neighbor
۲۴ Sanoja
۲۵ Gancarski
۲۶ Block-o-matic: A web page segmentation framework
۲۷ Ciarelli

عنوان "کاربرد یادگیری افزایشی چندبرچسبی در دسته‌بندی صفحات وب[۳]" منتشر کرده‌اند. پژوهشگران ذکر شده در این تحقیق با استفاده از یادگیری افزایشی چند برچسبی و هم‌چنین شبکه‌ی عصبی موفق شده‌اند دقت نتایج دسته‌بندی صفحات وب را بهبود بخشند. (Ciarelli, et al., 2014)

• راج[۴]، فرانسیس[۵] و بنادیت[۶] در سال ۲۰۱۶ مقاله‌ای انتشار داده‌اند تحت عنوان "تکنیک بهینه‌ی طبقه‌بندی صفحات وب براساس استخراج اطلاعات مفید و FA-NBC[۷]". آن‌ها در این مقاله با بهینه‌سازی الگوریتم‌های موجود قبلی و با استفاده از درخت تصمیم، الگوریتمی بر پایه‌ی الگوریتم نایو بیز ارائه کرده‌اند و مدعی شده‌اند که این الگوریتم از دیگر الگوریتم‌های طبقه‌بندی صفحات اینترنتی از جمله الگوریتم k-نزدیک‌ترین همسایه دقت بهتری دارد. (Raj, et al., 2016)

پژوهش‌هایی که به برخی از آن‌ها اشاره شد، اغلب از متن‌کاوی استفاده کرده‌اند. در زبان‌های انگلیسی، فرانسوی، عربی، اسپانیولی و امثالهم علاوه بر مدل‌های ساخته شده‌ای که برای متن‌کاوی به صورت آماده وجود دارد، داده‌های آموزش[۸]، کلمات زائد[۹] و دیگر مجموعه‌های مورد نیاز نیز در دسترس هستند؛ لکن در زبان فارسی تقریبا هیچ‌کدام از موارد ذکر شده در دسترس نیستند و لذا برای انجام پیش‌پردازش‌های مورد نیاز برای متن‌کاوی –از قبیل تکه‌تکه کردن[۱۰]، *stemming*، *lemmatization* و حذف کلمات زائد[۱۱]– نیازمند متناسب‌سازی آن مراحل برای زبان فارسی با تحقیق و پژوهش و کدنویسی بسیار هست. البته قابل ذکر است برخی از پژوهشگران از تکنیک‌های دیگری نیز در کنار متن‌کاوی استفاده نموده‌اند (مانند *Attardi, et al., 1999*) که در این تحقیق نیز محققان از این قبیل تکنیک‌ها در جای خود بهره برده‌اند.

از دهه‌ی گذشته تحقیقاتی در باب روند وب‌گردی کاربران انجام شده است. برخی از این تحقیقات فعالیت کاربران در قسمت‌های مختلف یک وب‌سایت را رصد و تحلیل کرده‌اند، برخی مسیر وب‌گردی کاربران از وب‌سایتی به دیگر وب‌سایت‌ها با استفاده از لینک‌ها را مورد توجه قرار داده‌اند و برخی از محققان نیز به تحلیل روند مراجعه‌ی کاربران به وب‌سایت‌های مختلف پرداخته‌اند. در ادامه فقط به عنوان نمونه به چند مقاله در این باب اشاره شده است:

• کولی[۱۲]، مباشر[۱۳] و سریواستاوا[۱۴] در سال ۱۹۹۹ مقاله‌ای منتشر کرده‌اند با عنوان "آماده‌سازی داده‌ها به منظور کاوش الگوهای وب‌گردی[۱۵]". در این پژوهش داده‌ها از سیاهه[۱۶]ی سرور جمع‌آوری شده و با استفاده از یک سری تکنیک آماده‌ی ورود بر مرحله‌ی داده‌کاوی می‌شوند. (Cooley, et al., 1999)

• ژو[۱۷] و لیو[۱۸] در سال ۲۰۱۰ طی مقاله‌ای با عنوان "تجزیه و تحلیل خوشه‌ای کاربران وب براساس الگوریتم k-means[۱۹]" کاربران وب را در جامعه‌ی نمونه‌ی خود براساس روند وب‌گردی‌شان خوشه‌بندی کرده‌اند. این دو محقق داده‌های وب‌گردی کاربران جامعه‌ی آماری خود را از به طور مستقیم از تاریخچه[۲۰]ی سیستم آن‌ها استخراج نموده‌اند و سپس با استفاده از الگوریتم k-means آن‌ها را خوشه‌بندی کرده‌اند. (Xu & Liu, 2010)

• ون[۲۱]، جانسون[۲۲]، ونگ[۲۳]، لی[۲۴] و یانگ[۲۵] در سال ۲۰۱۲ مشترکا مقاله‌ای ارائه‌کرده‌اند با عنوان "خوشه‌بندی کاربران وب با استفاده از شاخصه‌گذاری تصادفی با توابع وزنی[۲۶]". این پنج محقق در این مقاله فایل سیاهه‌ی وب‌گردی کاربران جامعه‌ی آماری خود را به صورت مستقیم گرفته و با

---

۱ Oliveira
۲ Salles
۳ Multi-label incremental learning applied to web page categorization.
۴ Raj
۵ Francis
۶ Benadit
۷ Optimal Web Page Classification Technique Based on Informative Content Extraction and FA-NBC (Firefly Algorithm based Naive Bayes Classifier)
۸ Train Data
۹ Stop Words
۱۰ Tokenizing
۱۱ Stop Words Removal

۱۲ Cooley
۱۳ Mobasher
۱۴ Srivastava
۱۵ Data preparation for mining world wide web browsing patterns
۱۶ log
۱۷ Xu
۱۸ Liu
۱۹ Web user clustering analysis based on KMeans algorithm
۲۰ history
۲۱ Wan
۲۲ Jonsson
۲۳ Wang
۲۴ Li
۲۵ Yang
۲۶ Web user clustering and Web prefetching using Random Indexing with weight functions

تعامل انسان و اطلاعات
جلد چهارم، شماره دوم، ۷۱-۸۷
http://hii.khu.ac.ir

رویکردی بر پایه‌ی مدل فضای برداری که شاخصه‌گذاری تصادفی[۱] نامیده می‌شود، کاربران وب را بر اساس رفتارشان در فضای وب خوشه‌بندی می‌کند. (Wan, et al., 2012)

• علی[۲] و الربیقی[۳] در سال ۲۰۱۶ مقاله‌ای منتشر کرده‌اند با عنوان "خوشه‌بندی کاربران وب براساس الگوریتم C-means فازی[۴]". نوآوری این دو محقق در این مقاله استفاده از منطق فازی برای انتساب کاربران وب به خوشه‌های تشخیص داده شده است، این‌چنین هر کاربر به یک نسبتی به هر کدام از خوشه‌ها نسبت داده می‌شود. (Ali & Alrabighi, 2016)

• آنیتا[۵] در سال ۲۰۱۶ طی مقاله‌ای با عنوان "یک الگوریتم خوشه‌بندی تجمیعی کارآمد برای شناسایی الگوهای وب‌گردی[۶]" الگوریتمی به منظور خوشه‌بندی کاربران وب ارائه داده است. قابل ذکر است وی در این مقاله داده‌های مورد نیازش را از طریق فایل‌های سیاهه‌ی وب‌گردی کاربران جمع‌آوری کرده است. (Anitha, 2016)

• دشماخ[۷] و ادهیا[۸] در سال ۲۰۱۶ مقاله‌ای تحت عنوان "مروری بر یافتن رفتار گردش کاربران با استفاده از الگوریتم‌های وب‌کاوی[۹]" ارائه کرده‌اند. این دو محقق در این مقاله ضمن مرور مقالات مرتبط پیشین، با استفاده از فایل سیاهه‌ی وب‌گردی کاربران، آن‌ها را خوشه‌بندی نموده‌اند. (Deshmukh & Adhiya, 2016)

• دهاراماراجان[۱۰] و دوریاناگاسوامی[۱۱] در سال ۲۰۱۶ مقاله‌ای تحت عنوان "کشف الگوهای تحلیل کاربران وب با استفاده از داده‌های سیاهه‌ی وب‌گردی آن‌ها و متخصصان سیاهه‌ی وب[۱۲]" منتشر کرده‌اند. این دو محقق در این تحقیق با استفاده از داده‌های سیاهه‌ی وب‌گردی کاربران وب که توسط ابزارهای آماده‌ای در دسترس قرار گرفته است، کاربران را خوشه‌بندی نموده و اطلاعاتی آماری از رفتار آن‌ها استخراج کرده‌اند. (Dharmarajan & Dorairangaswamy, 2016)

• سین[۱۳] و اسول[۱۴] در سال ۲۰۱۶ مقاله‌ای ارائه داده‌اند با عنوان "ارائه‌ی چارچوبی برای سیستم پیشنهاددهنده‌ی صفحات وب بر پایه‌ی وب‌کاوی معنایی[۱۵]". این دو در این مقاله یک چارچوب برای سیستم پیشنهاددهنده‌ی صفحات وب ارائه کرده‌اند که با استفاده از رفتار کاربر و دیگر کاربران، در هر لحظه به وی صفحه‌ای از وب برای مشاهده پیشنهاد می‌دهد. (Singh & Aswal, 2016)

• ژی[۱۶] و ونگ[۱۷] در سال ۲۰۱۶ مقاله‌ای تحت عنوان "پیشنهاد صفحه‌ی وب با خوشه‌بندی دوگانه: با توجه به رفتار کاربر و موضوع رابطه[۱۸]" ارائه کرده‌اند. این دو محقق در این تحقیق با ترکیب دو روش خوشه‌بندی از نقاط قوت هر دو سود برده‌اند. این دو ابتدا با استفاده از خوشه‌بندی مبتنی بر چگالی[۱۹] روی قسمتی از داده‌ها، تعداد و مرکز خوشه‌ها را پیدا کرده‌اند و در ادامه با استفاده از الگوریتم k-means خوشه‌بندی اصلی را انجام داده‌اند. (Xie & Wang, 2016)

پژوهش‌هایی که به در بالا بدان‌ها اشاره شد و دیگر پژوهش‌های مشابه در باب حل این مسئله، تقریبا تماما از داده‌های آماده یا سهل‌الوصول استفاده کرده‌اند –اغلب از فایل سیاهه‌ی وب‌گردی افراد یا تاریخچه‌ی وب‌گردی آن‌ها استفاده نموده‌اند و برخی نیز به اطلاعات وب‌گردی کاربران موجود در سمت سرورهای خدمات‌دهنده تکیه کرده‌اند– ولی در این پژوهش به دلیل اینکه یکی از اهداف پژوهش تسهیل تصمیم‌گیری در موضوع فیلترینگ است نیاز به داده‌ی استفاده یا عدم استفاده‌ی کاربر از فیلترشکن وجود دارد و درنتیجه استفاده از داده‌های آماده و در قالب داده‌های بالا ممکن نیست و درنتیجه جمع‌آوری داده‌ها خود نیاز به طراحی و پیاده‌سازی یک فرآیند یا سیستم خواهد داشت.

روش

این تحقیق از نظر هدف و جهت‌گیری یک تحقیق کاربردی

---

۱ Random Indexing
۲ Ali
۳ Alrabighi
۴ Web Users Clustering Based on Fuzzy C-MEANS
۵ Anitha
۶ An Efficient Agglomerative Clustering Algorithm for Web Navigation Pattern Identification
۷ Deshmukh
۸ Adhiya
۹ A Review on Finding Users Navigation Behavior Using Web Mining Algorithm
۱۰ Dharmarajan
۱۱ Dorairangaswamy
۱۲ Discovering User Pattern Analysis from Web Log Data using Weblog Expert
۱۳ Singh
۱۴ Aswal
۱۵ Towards a framework for web page recommendation system based on semantic web usage mining: A case study
۱۶ Xie
۱۷ Wang
۱۸ Web page recommendation via twofold clustering: considering user behavior and topic relation
۱۹ Density-based Clustering

و از نظر نحوه‌ی گردآوری داده‌ها یک تحقیق توصیفی-آزمایشی بوده، در قلمرو موضوعی جامعه‌شناسی و سیستم‌های اطلاعاتی قرار دارد و از بعد مکانی کاربران حاضر در ایران را هدف قرار داده است. قابل ذکر است که جامعه‌ی تحقیق این پژوهش کل کاربران اینترنتی در ایران بوده ولی جامعه‌ی آماری آن را کاربرانی تشکیل می‌دهند که افزونه۱ی ساخته شده را نصب کرده باشند.

در ادامه روش تحقیق را برمبنای الگوی کریسپ – دی ام[۲] (روش شناسی گام های داده کاوی مشترک بین صنایع) به اختصار توضیح می‌دهیم:

○ درک کسب‌وکار

قبل از تعریف عنوان و هدف پژوهش و در واقع در زمان طرح مسئله‌ی پژوهش، با کنکاش و پرسش پژوهشگران این تحقیق از افرادی از طیف‌های گوناگون جامعه، درک آن‌ها از فضای واقعی این حیطه افزایش یافته است.

○ درک داده

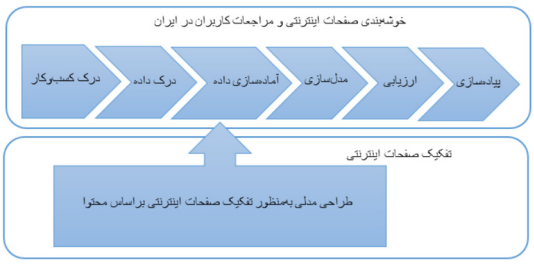
در این قسمت با هدف جمع‌آوری داده‌های وب‌گردی افراد نمونه، به طراحی، توسعه و انتشار افزونه‌های قابل نصب بر روی مرورگرهای مختلف پرداخته شده است. قابل‌ذکر است به منظور جلب نظر کاربران در استفاده از این افزونه‌ها، داشبوردی برای رصد وب‌گردی هر فرد به منظور استفاده‌ی وی در این افزونه‌ها تعبیه گردیده است.

○ آماده‌سازی داده

در این مرحله داده‌های جمع‌آوری شده در پایگاه داده‌ی تراکنشی استخراج شده، پاکسازی شده، پردازش‌های لازم به منظور استخراج اطلاعات دیگری از آن‌ها –از قبیل محتوای صفحه، کشور کاربر و رتبه در الکسا– انجام گرفته، تبدیلات لازم در قالب آن‌ها اعمال شده و در نهایت در پایگاه‌داده‌ی تحلیلی ذخیره می‌شوند.

○ مدل‌سازی

در این مرحله با استفاده از الگوریتم‌های داده‌کاوی مدلی به‌منظور دسته‌بندی صفحات اینترنتی و هم‌چنین مدلی به‌منظور دسته‌بندی نوع مراجعات کاربران به وب‌سایت‌ها طراحی می‌شود.



○ ارزیابی

در انتها مدل ساخته شده، ارزیابی شده و با تفسیر دسته‌ها، اعتبار آن را نشان داده شده است.

○ پیاده‌سازی

و در آخرین مرحله داده‌های جمع‌آوری‌شده را پردازش شده و نهایتا به کمک مدل‌های ساخته‌شده، صفحات اینترنتی مورد رجوع کاربران و نوع مراجعاتشان دسته‌بندی شده و مورد تفسیر و تحلیل قرار گرفته است.

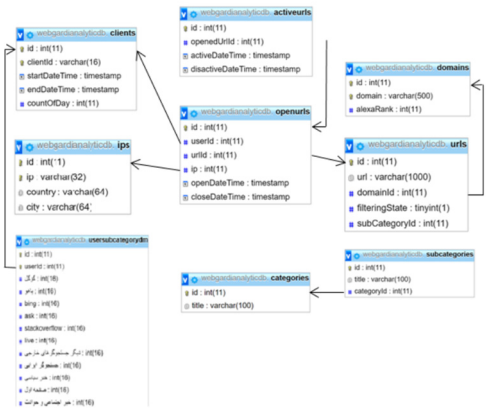
جمع‌آوری داده

با توجه به اینکه یکی از اهداف این تحقیق تسهیل تصمیم‌گیری در مقوله‌ی فیلترینگ در کشور بوده است، بایستی اطلاعات مورد نیاز مستقیما از سمت کاربران جمع‌آوری می‌شد. از طرفی با توجه به جو امنیتی موجود در ذهن هم‌وطنان عزیز، این نیاز احساس شد که بایستی یک خدمتی به آن‌ها ارائه شود تا آن‌ها اطلاعات مورد نیاز ما را در اختیار پژوهشگران این تحقیق قرار دهند. لذا بدین منظور دو نسخه افزونه‌ی قابل نصب بر روی مرورگرهای کروم و فایرفاکس تولید شد که کاربران اینترنتی با نصب آن‌ها بر روی مرورگر خود می‌توانستند به صورت لحظه‌به‌لحظه وضعیت وب‌گردی خود را در قالب یک داشبورد[۳] حاوی نمودارها و شاخص[۴]های تعریف‌شده ببینند. با انتشار این افزونه‌ها در *Mozilla add-ons* و *Google Webstore* و هم‌چنین شبکه‌های اجتماعی _از قبیل تلگرام، گوگل‌پلاس، اینستاگرام و آپارات_ در کمتر از دو هفته تعداد نصب این افزونه‌ها از مرز ۱۵۰ نصب رد شد. قابل ذکر است که این افزونه بدین صورت عمل می‌کرد که در زمان نصب، قوانین و حقوق مولفین و کاربر را ذکر کرده و با تایید کاربر، کد یکتایی به وی نسبت می‌داد. در ادامه با هر اقدام کاربر در مرورگر _شامل باز کردن تب جدید، بستن تب و رفتن به آدرس جدید_ اطلاعات این اقدام سمت سرور ما ارسال شده و در پایگاه‌داده‌ی عملیاتی[۵] طراحی شده ذخیره می‌شد. شمای کلی پایگاه‌داده‌ی عملیاتی ما به صورت شکل ۱ است. در داشبورد طراحی‌شده از برخی اطلاعات مانند کشور کاربر و دسته‌ی محتوایی صفحات اینترنتی استفاده شده است که این

[۳] dashboard
[۴] KPI
[۵] ODB

تعامل انسان و اطلاعات
جلد چهارم، شماره دوم، ۷۱-۸۷
http://hii.khu.ac.ir

شکل ۱. شمای کلی پایگاه‌داده‌ی عملیاتی



شکل ۲. شمای کلی انباره‌ی داده

اطلاعات از پردازش روی داده‌های جمع‌آوری‌شده‌ی موجود در پایگاه‌داده‌ی عملیاتی حاصل می‌شدند. از طرفی به علت تنوع و پیچیدگی برخی از نمودارهای داشبورد طراحی‌شده، سرعت بارگذاری داشبورد شخصی کاربران پایین بود. به این دو دلیل یک انباره‌ی داده[۱] طراحی شده که یک مکعب[۲] را نیز شامل می‌شود. شمای کلی انباره‌ی داده‌ی طراحی‌شده به صورت شکل ۱ است.

همچنین برنامه‌ای در نقش یک *ETL*[۳] نوشته شده که هر پنج دقیقه یک‌مرتبه اجرا شده، داده‌های جدید پایگاه‌داده‌ی عملیاتی را خوانده، اطلاعاتی مانند کشور کاربر، دسته‌ی محتوایی صفحه‌ی اینترنتی، رتبه‌ی الکسای سایت مورد رجوع و مدت زمان باز بودن صفحه را با پردازش داده‌های پایگاه‌داده‌ی عملیاتی و رجوع به پایگاه‌داده‌ها و وب‌سرویس‌های آماده شده‌ی موجود در فضای وب، به دست آورده و در انباره‌ی داده ذخیره می‌کند.

تفکیک محتوایی صفحات اینترنتی

همانطور که اشاره شد برای تولید داشبورد شخصی طراحی‌شده و البته برای خوشه‌بندی نهایی، نیاز به تفکیک صفحات اینترنتی بر اساس محتوای آن‌ها وجود داشته است. این قسمت از تحقیق که خود به‌عنوان یک پژوهش مجزا می‌تواند به صورت عمیق‌تر مورد پژوهش قرار گیرد به این صورت انجام شد که ابتدا ۱۴ دسته و حدود ۱۰۰ زیردسته‌ی محتوایی مختلف توسط پژوهشگران و با مشورت خبرگان این حوزه تعریف شده است.

در گام اول یک دسته‌بندی دو سطحی برای سایت‌ها در نظر گرفته شد. سپس ۱۰۰ سایت پربازدید ایران بر اساس رتبه‌بندی الکسا انتخاب شده و در صف دسته‌بندی قرار گرفتند. همچنین عناوین مرتبط با هریک از زیردسته‌ها در جستجوگرهای گوگل و یاهو جستجو شدند و نتایج حاصل در ۲ الی ۵ صفحه‌ی ابتدایی به صف شناسایی اضافه شدند. همچنین سایت‌های پربازدید خارجی براساس رتبه‌بندی الکسا که فیلتر بوده و به این علت ممکن است در لیست پربازدیدهای ایران قرار نگرفته‌باشند نیز به صف مذکور اضافه شدند. حال بایستی وب‌سایت‌های انتخاب شده دسته‌بندی شوند. بدین منظور به بررسی تک‌تک این وب‌سایت‌ها پرداخته شده است.

برخی از آن‌ها به طور کامل در یک زیردسته قرار می‌گرفتند. پس اگر آدرس صفحه‌ی مورد نظر متعلق به این مجموعه دامنه باشد، به آسانی می‌توان زیردسته و به‌تبع آن، دسته‌ی آن صفحه را مشخص نمود. این مجموعه بیش از ۱۳۶۰ دامنه را شامل می‌شود.

برخی دیگر از وب‌سایت‌ها چند زیردسته را شامل می‌شدند که برخی از آن‌ها از طریق آدرس قابل تفکیک به زیردسته‌های مختلف هستند. در صورتی که صفحه‌ی مورد نظر در مرحله‌ی قبل دسته‌بندی نشده باشد و جزو این مجموعه باشد در این مرحله دسته‌بندی خواهد شد. این مجموعه شامل بیش از ۴۵ دامنه می‌باشد.

پس از این دو مرحله نیز برخی از صفحات هنوز دسته‌بندی نشده‌اند. با بررسی تک‌تک آن‌ها مشاهده می‌شود برخی از آن‌ها را می‌توان با استخراج قسمت مشخصی از صفحه دسته‌بندی نمود. لذا به‌ازای هر یک از دامنه‌های موجود در این دامنه یک تابع می‌نویسیم که با خزش[۴] روی صفحه‌ی مورد نظر، قسمت مشخص را استخراج نموده و طبق آن صفحه‌ی مورد نظر را دسته‌بندی کند. این مجموعه شامل ۲۴ دامنه می‌باشد.

در انتها با بررسی انجام شده روی داده‌هایی که تا به‌حال جمع‌آوری شده‌اند، مشاهده شد که با انجام این سه مرحله،

---

[۱] Data warehouse

[۲] cube

[۳] Extract, Transform, Load

[۴] Crawl

بیش از ۹۲ درصد صفحات بازدید شده، دسته‌بندی شده‌اند.در گام آخر با استفاده از ساده‌ترین و سبک‌ترین الگوریتم دسته‌بندی متون یعنی الگوریتم مدل فضای برداری به دسته‌بندی صفحاتی که در سه مرحله‌ی فوق دسته‌بندی نشدند، پرداخته شده و این‌چنین پژوهشگران این تحقیق توانسته‌اند تمامی صفحات اینترنتی را از نظر محتوایی دسته‌بندی کنند.

خوشه‌بندی[1] صفحات مورد رجوع

در خوشه‌بندی صفحات اینترنتی مورد رجوع کاربران به سه ویژگی توجه شده است:

- دسته‌ی محتوایی: در بخش قبل با استفاده از یک سری اعمال حریصانه[2] و نهایتا تکنیک متن‌کاوی، صفحات اینترنتی به حدود ۱۰۰ زیردسته و ۱۴ دسته نسبت داده شده است. حال در این‌جا تصمیم گرفته شده فقط از دسته‌های محتوایی مختلف به عنوان ورودی الگوریتم خوشه‌بندی استفاده شود.
- لگاریتم رتبه‌ی الکسای صفحه: در برنامه‌ای که به منظور ETL نوشته شده است، رتبه‌ی هر دامنه را با خزش روی سایت الکسا به دست آورده و ثبت گردیده است. منتها با توجه به اینکه فاصله‌ها در رتبه‌های بالاتر با معنادارتر است (به عنوان مثال فاصله‌ی رتبه‌های ۱ با ۱۰۰۰ خیلی بامعنی‌تر از فاصله‌ی رتبه‌های صد هزار با صد و یک هزار می‌باشد) تصمیم گرفته شده به جای دخیل کردن رتبه‌ی الکسای هر دامنه، لگاریتم رتبه در خوشه‌بندی دخالت داده شود.
- وضعیت فیلتر بودن صفحه: به منظور استخراج وضعیت فیلتر بودن یا نبودن صفحات اینترنتی، یک خزنده نوشته شده تا روی تمامی ۱۳۱۱۴۴ صفحه‌ی اینترنتی ثبت شده در جدول urls خزش کند و وضعیت فیلتر بودن یا نبودن آن را تشخیص داده و در پایگاه داده ثبت کند.

با توجه به این‌که هم داده‌ی از نوع عددی داریم و هم داده‌ی از نوع دسته‌ای، نمی‌توان از الگوریتم‌های معمول خوشه‌بندی استفاده نمود. با جستجو در مقالات علمی سه راه‌حل برای این مشکل پیدا شده است:

- دسته‌ای کردن داده‌های عددی و استفاده از الگوریتم k-modes
- استفاده از الگوریتم k-prototype
- استفاده از الگوریتم PCL-OC

با توجه به مزایا و معایب هر کدام از راه‌حل‌های فوق (Huang, 1998. Huang, 1997 & Cheung & Jia, 2013.) و البته مقایسه‌ی نتایج خوشه‌بندی براساس راه‌حل‌های دوم و سوم توس پژوهشگران این تحقیق، راه‌حل سوم را برای خوشه‌بندی برگزیده شد.

خوشه‌بندی نوع مراجعات

در خوشه‌بندی نوع مراجعات کاربران به سه ویژگی توجه شده است:

- زمان مراجعه: زمان مراجعه‌ی کاربر به صفحه که یک زمان در ۲۴ ساعت شبانه‌روز بود در ۴ طبقه (۱۲ بامداد تا ۶ صبح، ۶ صبح تا ۱۲ ظهر، ۱۲ ظهر تا ۶ عصر و ۶ عصر تا ۱۲ بامداد) خلاصه شده است. بدین ترتیب این فیلد عددی به فیلدی دسته‌ای تعدیل شد.
- مدت زمان رجوع: با نگاهی به مدت زمان رجوع که برحسب ثانیه است، پژوهشگران متوجه شدند که گاه تا چند ده هزار ثانیه ثبت شده است و این‌چنین نیاز به شناسایی و حذف داده‌های خارج از محدوده[3] حس شد. به منظور شناسایی داده‌های خارج از محدوده در این فیلد، تصمیم به استفاده از تکنیک خوشه‌بندی گرفته شد. از طرفی برای خوشه‌بندی الگوریتم‌های مختلفی وجود دارد، پژوهشگران در این‌جا برای انتخاب بهترین خوشه‌بندی از معیار Silhouette استفاده کرده‌اند؛ بدین منظور داده‌های مدت زمان رجوع را توسط الگوریتم‌های MeanShift، KMeans، DBSCAN و Ward hierarchical clustering و با ورودی‌های مختلف خوشه‌بندی نموده و مقدار معیار Silhouette را برای نتایج این خوشه‌بندی‌ها محاسبه کرده‌اند. در انتها بیشترین مقدار معیار Silhouette (۰,۹۸) مربوط به خوشه‌بندی با استفاده از الگوریتم KMeans و با تعداد خوشه‌ی مساوی ۲، به دست آمد. این خوشه‌بندی ۱۴۲۰۳۶ داده را در دو خوشه تقسیم کرد: ۱۴۱۵۱۶ داده در خوشه‌ی اول و ۵۲۰ داده در خوشه‌ی دوم. و بدین صورت ۵۲۰ داده را به عنوان داده‌های خارج از محدوده شناسایی کرده و از داده‌های مورد نظر حذف گردیده است.
- استفاده یا عدم استفاده‌ی کاربر از فیلترشکن: این ویژگی را بر اساس کشور کاربر که از IP وی استخراج شده بود، تشخیص داده شده است.

در این خوشه‌بندی نیز با توجه به این‌که هم داده‌ی از نوع

[1] Clustering
[2] Greedy
[3] Outliers

جدول ۱. خوشه‌های صفحات اینترنتی

شناسه‌ی خوشه	عنوان پیشنهادی	تعریف
۱	ارتباطی فیلترنشده‌ی پربازدید	ابزارهای ارتباطی فیلترنشده با رتبه‌ی الکسای ۱۴ و ۱۶ (توئیتر و اینستاگرام)
۲	خبری فیلترنشده‌ی کم‌بازدید	سایت‌های خبری فیلترنشده‌ی کم‌بازدید (رتبه‌ی الکسای بالای هزار)
۳	ارتباطی فیلترنشده‌ی کم‌بازدید	ابزارهای ارتباطی فیلترنشده‌ی کم‌بازدید (رتبه‌ی الکسای بالای ۴۰۰)
۴	سایت دانلود فیلترنشده	سایت‌های دانلود فایل، بازی، نرم‌افزار، موسیقی، فیلم و ... فیلترنشده
۵	علمی فیلترنشده	صفحات علمی-آموزشی فیلترنشده
۶	خرید اینترنتی	سایت‌های خرید اینترنتی
۷	سرگرمی فیلترنشده	سایت‌های سرگرمی (بازی آنلاین، داستان، لطیفه، اخبار زرد، فال و تعبیر خواب و ...) فیلترنشده
۸	خبری فیلترنشده‌ی پربازدید	سایت‌های خبری فیلترنشده‌ی پربازدید (رتبه‌ی الکسای زیر هزار)
۹	وبلاگ فیلترنشده	وبلاگ‌های فیلترنشده
۱۰	پورتال پربازدید	پورتال و شبکه‌های اطلاع‌رسانی پربازدید (رتبه‌ی الکسای زیر ۹۵۰۰)
۱۱	ارتباطی فیلترنشده با تعداد بازدید متوسط	ابزارهای ارتباطی فیلترنشده با رتبه‌ی الکسای بین ۲۰ و ۴۰۰
۱۲	خبری فیلترشده	سایت‌های خبری فیلترشده
۱۳	سرگرمی فیلترشده	سایت‌های سرگرمی (بازی آنلاین، داستان، لطیفه، اخبار زرد، فال و تعبیر خواب و ...) فیلترشده
۱۴	وبلاگ فیلترشده	وبلاگ‌های فیلترشده
۱۵	علمی فیلترشده	صفحات علمی-آموزشی فیلترشده
۱۶	عمومی فیلترنشده‌ی کم‌بازدید	صفحات عمومی (مذهبی، فرهنگی، روانشناسی، آشپزی، آرایشی، گردشگری و ...) فیلترنشده‌ی کم‌بازدید (رتبه‌ی الکسای بالای ده میلیون)
۱۷	ارتباطی فیلترنشده‌ی خیلی پربازدید	ابزارهای ارتباطی فیلترنشده با رتبه‌ی الکسای ۱ و ۵ (ابزارهای ارتباطی گوگل از قبیل جی‌میل و گوگل پلاس و ابزارهای ارتباطی یاهو از قبیل ایمیل یاهو)
۱۸	سایت دانلود فیلترشده	سایت‌های دانلود فایل، بازی، نرم‌افزار، موسیقی، فیلم و ... فیلترشده
۱۹	ابزار کاربردی فیلترشده	ابزار کاربردی (ابزار مدیریت روزمره، ترجمه، خدمات سرور و ...) فیلترشده
۲۰	ابزار کاربردی فیلترنشده‌ی پربازدید	ابزار کاربردی (ابزار مدیریت روزمره، ترجمه، خدمات سرور و ...) فیلترنشده‌ی پربازدید (رتبه‌ی الکسای زیر ده هزار)
۲۱	موتور جستجو	موتورهای جستجو
۲۲	عمومی فیلترنشده‌ی پربازدید	صفحات عمومی (مذهبی، فرهنگی، روانشناسی، آشپزی، آرایشی، گردشگری و ...) فیلترنشده‌ی پربازدید (رتبه‌ی الکسای زیر ده میلیون)
۲۳	آگهی استخدام	آگهی استخدام
۲۴	ابزار کاربردی فیلترنشده‌ی کم‌بازدید	ابزار کاربردی (ابزار مدیریت روزمره، ترجمه، خدمات سرور و ...) فیلترنشده‌ی کم‌بازدید (رتبه‌ی الکسای بالای ده هزار)
۲۵	ارتباطی فیلترشده	ابزارهای ارتباطی فیلترشده
۲۶	پورتال کم‌بازدید	پورتال و شبکه‌های اطلاع‌رسانی کم‌بازدید (رتبه‌ی الکسای بالای ۹۵۰۰)
۲۷	عمومی فیلترشده	صفحات عمومی (مذهبی، فرهنگی، روانشناسی، آشپزی، آرایشی، گردشگری و ...) فیلترشده
۲۸	غیراخلاقی	صفحات غیراخلاقی
۲۹	تبلیغات	تبلیغات اینترنتی

عددی وجود دارد و هم داده‌ی از نوع دسته‌ای، از الگوریتم PCL-OC استفاده شده است.

یافته‌ها

انجام خوشه‌بندی صفحات مورد رجوع کاربران، صفحات اینترنتی را در ۲۹ خوشه قرار داد. در جدول ۱ این خوشه‌ها تفسیر شده‌اند.

انجام خوشه‌بندی مراجعات کاربران، مراجعات کاربران را در ۱۱ خوشه قرار داد. در جدول ۲ این خوشه‌ها تفسیر پس از خوشه‌بندی موفقیت‌آمیز صفحات مورد رجوع و نوع مراجعات کاربران و تحلیل و تعریف دقیق هر خوشه، گزارش‌های متنوعی می‌توان آماده کرد که در ادامه فقط به عنوان نمونه، چند گزارش ارائه می‌شود.

تحلیل رفتاری هر یک از کاربران: با توجه به نوع مراجعات

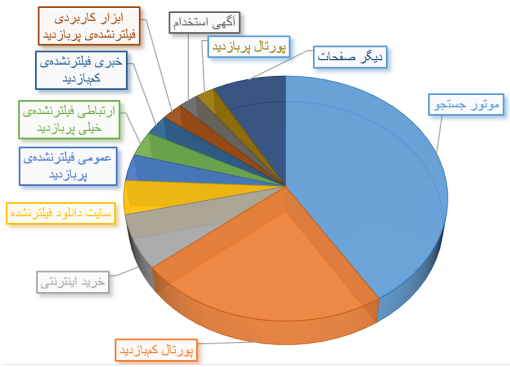
جدول ۲. خوشه‌های مراجعات کاربران

شناسه‌ی خوشه	عنوان پیشنهادی	تعریف
۱	بازدید عصرگاهی چندثانیه‌ای بدون استفاده از فیلترشکن	رجوع به یک صفحه‌ی اینترنتی در بازه‌ی زمانی۱۲ ظهر تا ۶ عصر بدون استفاده از فیلترشکن و کمتر از یک دقیقه ماندن در این صفحه
۲	بازدید صبح‌گاهی با استفاده از فیلترشکن و بازدید صبح‌گاهی طولانی‌مدت بدون استفاده از فیلترشکن	رجوع به یک صفحه‌ی اینترنتی در بازه‌ی زمانی ۶ صبح تا ۱۲ ظهر با استفاده از فیلترشکن و رجوع به یک صفحه‌ی اینترنتی در بازه‌ی زمانی۶ صبح تا ۱۲ ظهر بدون استفاده از فیلترشکن و بیشتر از یک ساعت ماندن در این صفحه
۳	بازدید صبح‌گاهی چند دقیقه‌ای بدون استفاده از فیلترشکن	رجوع به یک صفحه‌ی اینترنتی در بازه‌ی زمانی ۶ صبح تا ۱۲ ظهر بدون استفاده از فیلترشکن و بیشتر از ۶ دقیقه ماندن در این صفحه
۴	بازدید سحر، صبح و عصرگاهی با استفاده از فیلترشکن	رجوع به یک صفحه‌ی اینترنتی در بازه‌ی زمانی ۱۲ بامداد تا ۶ عصر با استفاده از فیلترشکن
۵	بازدید سحرگاهی بدون استفاده از فیلترشکن	رجوع به یک صفحه‌ی اینترنتی در بازه‌ی زمانی ۱۲ بامداد تا ۶ صبح بدون استفاده از فیلترشکن
۶	بازدید عصرگاهی دو دقیقه‌ای بدون استفاده از فیلترشکن	رجوع به یک صفحه‌ی اینترنتی در بازه‌ی زمانی۱۲ ظهر تا ۶ عصر بدون استفاده از فیلترشکن و حدود دو دقیقه ماندن در این صفحه
۷	بازدید شبانگاهی چند ساعته بدون استفاده از فیلترشکن	رجوع به یک صفحه‌ی اینترنتی در بازه‌ی زمانی ۶ عصر تا ۱۲ بامداد بدون استفاده از فیلترشکن و بیشتر از صد دقیقه ماندن در این صفحه
۸	بازدید شبانگاهی چندین دقیقه‌ای بدون استفاده از فیلترشکن	رجوع به یک صفحه‌ی اینترنتی در بازه‌ی زمانی ۶ عصر تا ۱۲ بامداد بدون استفاده از فیلترشکن و بین سه تا صد دقیقه ماندن در این صفحه
۹	بازدید شبانگاهی یکی دو دقیقه‌ای بدون استفاده از فیلترشکن	رجوع به یک صفحه‌ی اینترنتی در بازه‌ی زمانی ۶ عصر تا ۱۲ بامداد بدون استفاده از فیلترشکن و کمتر از دو دقیقه ماندن در این صفحه
۱۰	بازدید صبح‌گاهی کوتاه‌مدت بدون استفاده از فیلترشکن	رجوع به یک صفحه‌ی اینترنتی در بازه‌ی زمانی ۶ صبح تا ۱۲ ظهر بدون استفاده از فیلترشکن و کمتر از ۶ دقیقه ماندن در این صفحه
۱۱	بازدید عصرگاهی طولانی‌مدت بدون استفاده از فیلترشکن	رجوع به یک صفحه‌ی اینترنتی در بازه‌ی زمانی۱۲ ظهر تا ۶ عصر بدون استفاده از فیلترشکن و بیشتر از سه دقیقه ماندن در این صفحه

هر کاربر و نسبت صفحات مورد رجوع وی، می‌توان رفتار وب‌گردی وی را تحلیل و تفسیر کرد. در ادامه فقط به عنوان نمونه، تحلیل رفتاری یک کاربر ارائه شده است.

در ابتدا در قالب یک نمودار دایره‌ای، نسبت مراجعه‌ی این کاربر به دسته‌های مختلف صفحات نشان داده شده است.

با توجه به نمودار ۳ مشخص است که این کاربر بیشتر مدت وب‌گردی خود را به جستجو پرداخته و همچنین در وهله‌ی



شکل ۳. نسبت مراجعه‌ی کاربر مورد نظر به دسته‌های مختلف صفحات اینترنتی

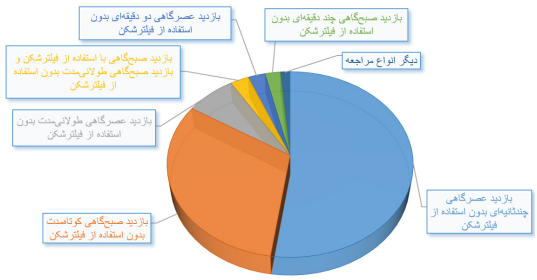
بعدی در پورتال (یا پورتال‌هایی) کم‌بازدید گردش کرده است. در ادامه باز هم در قالب یک نمودار دایره‌ای، انواع گوناگون مراجعات این کاربر نشان داده شده است:

همانطور که در نمودار دایره‌ای ۴ مشهود است، این کاربر بیشتر در بازه‌ی زمانی ظهر تا عصر و در وهله‌ی بعدی در بازه‌ی صبح تا ظهر به وب‌گردی پرداخته است. از طرفی اغلب مدت بسیار کوتاهی را در هر صفحه‌ی اینترنتی سپری کرده است. در ادامه به منظور بررسی دقیق‌تر ساعات وب‌گردی این کاربر، تصویر ۵ را تشکیل می‌دهیم.

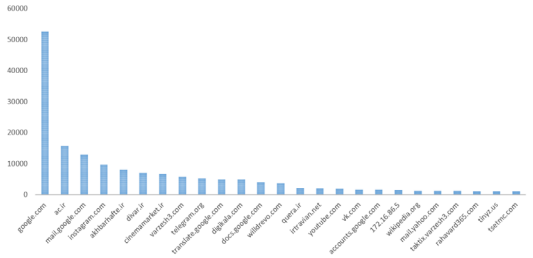
با توجه به میزان فعالیت کاربر در ساعات مختلف شبانه‌روز، می‌توان حدس زد که این کاربر در واقع یک سیستم در یک محیط اداری و یا دانشگاهی می‌باشد که افزونه‌ی ما روی آن نصب شده است. در انتها برای بررسی این فرضیه، وب‌سایت‌های پربازدید این کاربر را استخراج کرده و در قالب یک نمودار میله‌ای در تصویر ۶ نشان می‌دهیم.

با توجه به ۱۵ وب‌سایت پربازدید توسط این کاربر که در نمودار بالا قابل مشاهده است، می‌توان این حدس را تقویت نمود که این کاربر در واقع یک سیستم در یک محیط

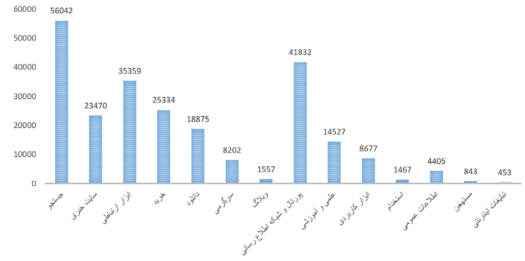
تعامل انسان و اطلاعات
جلد چهارم، شماره دوم، ۷۱-۸۷
http://hii.khu.ac.ir



شکل ۶. نسبت انواع گوناگون مراجعات کاربر مورد نظر



شکل ۷. وب‌سایت‌های پربازدید



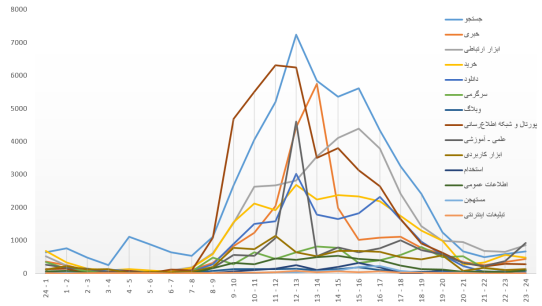
شکل ۸. میزان استفاده‌ی کاربران ایرانی از محتواهای مختلف در فضای

دانشگاهی است.

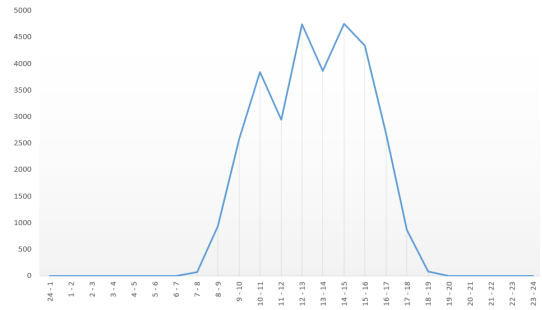
وب‌گردی کاربران ایرانی از نگاه کلان: برای ایجاد یک نگاه کلان دقیق، نیاز به داده‌های وسیعی است، از طرفی این پروژه یک پروژه‌ی پژوهشی است؛ فلذا در این قسمت ما با توجه به داده‌های محدود جمع‌آوری شده، سعی در ایجاد یک نگاه کلان به وضعیت وب‌گردی کاربران ایرانی می‌پردازیم؛ ولی باید توجه داشت که برای تصمیم‌گیری‌های کلان در مورد برخورد و تعامل با فضای مجازی و بالاخص وب، بایستی محدوده‌ی وسیع‌تری از داده‌ها را جمع‌آوری نمود و برپایه‌ی آن‌ها تصمیم‌گیری کرد. در ادامه در چند بخش و از چند وجه مختلف به وضعیت وب‌گردی کاربران ایرانی از نگاه کلان خواهیم پرداخت:

- وب‌سایت‌های پربازدید

شاید در نگاه اول به نظر آید که وب‌سایت‌های پربازدید توسط کاربران ایرانی را می‌توان از منابعی عمومی —مثلا سایت الکسا– به صورت دقیق‌تر به دست آورد؛ لکن با توجه به اینکه بخشی از کاربران از داخل ایران ولی با IP هایی



شکل ۴. نرخ مراجعه‌ی کاربران ایرانی به محتواهای مختلف در ساعات مختلف شبانه‌روز



شکل ۵. نمودار زمانی وبگردی کاربر مورد نظر در ساعات شبانه‌روز

خارج از کشور —با استفاده از فیلترشکن و VPN– به وب‌گردی می‌پردازند، دقت آن منابع کاهش می‌یابد. ما ادعا می‌کنیم که با سیستم طراحی کرده، دقتی به مراتب بالاتر از آن منابع را در این قسمت ارائه خواهیم داد. در تصویر ۷، ۲۵ وب‌سایت پربازدید کاربران تحت پوشش خود را در قالب یک نمودار میله‌ای ارائه کرده‌ایم.

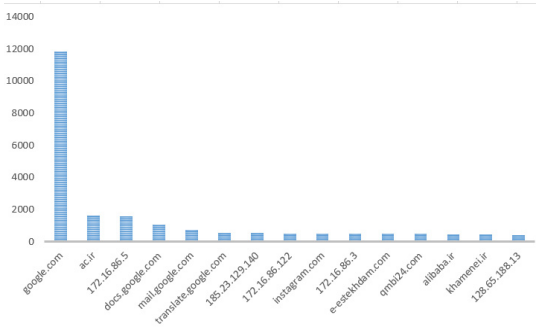
- میزان استفاده از محتواهای مختلف در فضای وب
- نرخ وب‌گردی در ساعات مختلف شبانه‌روز

با توجه به محدودیت کاربران در این پژوهش، به اطلاعات این بخش نمی‌توان زیاد اتکا نمود؛ با این‌حال با توسعه‌ی طیف کاربران استفاده‌کننده از این افزونه، این بخش اطلاعات مهمی را در اختیار ما قرار خواهد گذاشت که در تصمیم‌گیری برای بهبود سبک زندگی مردم کشور، بسیار مهم و راهگشا خواهد بود.

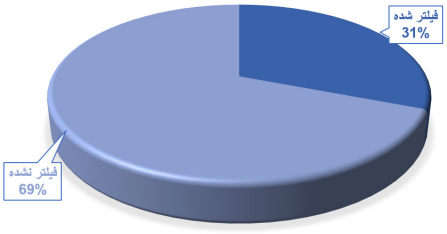
- نرخ مراجعه به محتواهای مختلف در ساعات شبانه‌روز

فیلترینگ و فیلترشکن در فضای وب

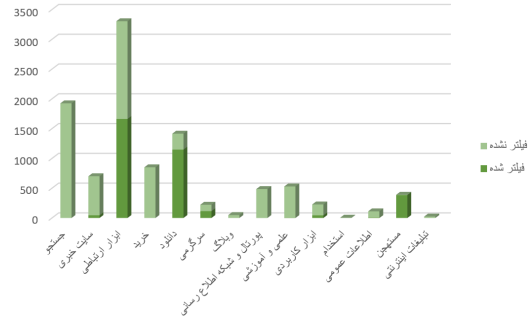
یکی از تصمیم‌گیری‌های کلان این روزهای کشور حول مبحث فیلترینگ می‌باشد؛ از این‌که فیلترینگ به چه صورتی و با چه شدتی انجام شود گرفته تا بررسی نتایج فیلترهای صورت گرفته، تا تصمیم‌گیری در باب برخورد یا توسعه‌ی فیلترشکن‌هایی با قابلیت رفع فیلتر صفحات خاص تا ... . با توجه به چند بعدی بودن این مبحث برای انجام تصمیم‌گیری



شکل ۱۲. وب‌سایت‌های پربازدید توسط کاربر مورد نظر



شکل ۱۳. میزان رجوع به صفحات فیلترشده و صفحات فیلترنشده در استفاده از فیلترشکن

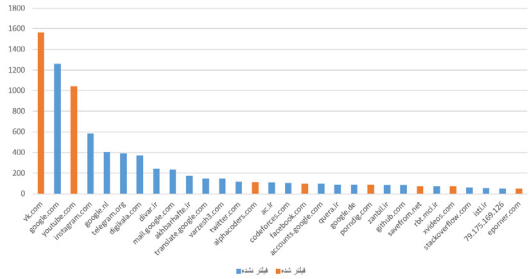


شکل ۱۴. میزان استفاده‌ی کاربران ایرانی وب از فیلترشکن در محتواهای مختلف

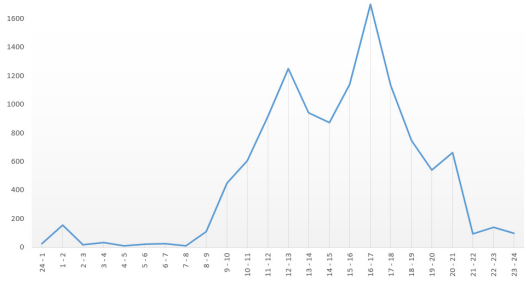
صحیح در این‌باره، نیاز به اطلاعاتی واقعی و کافی می‌باشد. در این راستا ما سعی کرده‌ایم یکی از خروجی‌های سیستم طراحی‌شده، گزارشاتی برای کمک به مسئولین این امر در کشور باشد. به‌عنوان نمونه چند بخش از گزارش پیشنهادی در ادامه ذکر می‌شود.

- وب‌سایت‌های پربازدید کاربران ایرانی با استفاده از فیلترشکن

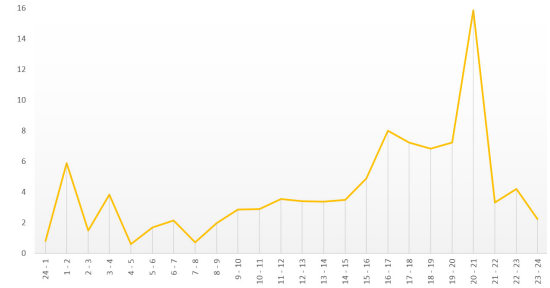
برای شروع ترجیح دادیم تا ببینیم کاربران ایرانی فضای وب، از فیلترشکن برای رجوع به چه وب‌سایت‌هایی استفاده می‌کنند. در نمودار میله‌ای زیر، به ۳۰ وب‌سایتی که کاربران



شکل ۹. وب‌سایت‌های فیلترشده و فیلترنشده‌ی پربازدید با استفاده از فیلترشکن



شکل ۱۰. میزان استفاده از فیلترشکن در ساعات مختلف شبانه‌روز



شکل ۱۱. درصد استفاده‌ی کاربران ایرانی وب از فیلترشکن در ساعات مختلف شبانه‌روز

ایرانی به دفعات بیشتری با استفاده از فیلترشکن به آن‌ها رجوع کرده‌اند، اشاره کرده‌ایم. قابل ذکر است که وب‌سایت‌های فیلترشده با رنگ متفاوت نسبت به وب‌سایت‌های فیلترنشده متمایز گشته‌اند.

- رفع فیلتر یا دور زدن تحریم؟

یکی از نکاتی که مسئولین مربوطه را نسبت به برخورد با فیلترشکن دچار شک کرده و حتی آن‌ها را ترغیب به توسعه‌ی فیلترشکن‌هایی خاص می‌کند، استفاده از فیلترشکن برای دور زدن تحریم‌هایی است که کاربران اینترنتی ما با آن مواجهند. برای کمک به آن‌ها در تصمیم‌گیری‌های بهتر و محکم‌تر، نیاز است تا بررسی کنیم که چه میزان استفاده از فیلترشکن برای رفع فیلتر بوده و چه میزان آن برای دور زدن تحریم‌های ظالمانه. در ادامه میزان استفاده از فیلترشکن در مراجعه به وب‌سایت‌های فیلترشده را نسبت به مراجعه به

تعامل انسان و اطلاعات
جلد چهارم، شماره دوم، ۷۱-۸۷
http://hii.khu.ac.ir

دیگر وب‌سایت‌ها در قالب یک نمودار دایره‌ای نشان می‌دهیم.

همان‌طور که در نمودار بالا مشخص است، تنها ۳۱ درصد از استفاده از فیلترشکن به‌منظور رفع فیلتر بوده است. البته باید توجه کرد که لزوما تمام ۶۹ درصد باقی‌مانده متعلق به استفاده به منظور دور زدن تحریم‌ها نبوده و قسمی از این میزان به عادت کاربران به استفاده از فیلترشکن برمی‌گردد!

در ادامه نیاز دیدیم که برای فهم بهتر فضای استفاده از فیلترشکن، نگاهی بیندازیم به استفاده از فیلترشکن برای مراجعه به دسته‌های محتوایی مختلف؛ بدین منظور نمودار زیر را تهیه کردیم. قابل توجه است که در این نمودار نیز صفحات فیلترشده و صفحات فیلترنشده با کمک رنگ از یکدیگر متمایز شده‌اند.

همانطور که از نمودار فوق مشخص است بیشتر استفاده از فیلترشکن برای رجوع به ابزارهای ارتباطی، موتورهای جستجو و دانلود بوده و صفحات علمی-آموزشی و ابزارهای کاربردی در مراتب پایین‌تری قرار دارند. همچنین قابل توجه است که بیشترین استفاده از فیلترشکن برای رفع فیلتر، مربوط به رفع فیلتر ابزارهای ارتباطی –از جمله فیسبوک–، صفحات دانلود –غالبا دانلود فیلم– و صفحاتی با محتوای مستهجن می‌باشد.

ساعات اوج استفاده از فیلترشکن: از دیگر اطلاعاتی که می‌توان در باب استفاده از فیلترشکن ارائه داد، نرخ استفاده از فیلترشکن در ساعات مختلف شبانه‌روز می‌باشد. در تصویر ۱۴ این اطلاعات در قالب نموداری آورده شده است.

با توجه به طیف محدود کاربران این پژوهش و عدم توزیع یکنواخت مراجعات در طول ۲۴ ساعت شبانه‌روز، در کنار نمودار بالا نموداری دیگر تهیه کرده‌ایم که درصد استفاده از فیلترشکن را در هر ساعت نشان می‌دهد.

بحث و نتیجه گیری

این تحقیق نسبت به پژوهش‌های مشابه پیشین چندین وجه نو دارد که در ادامه به چند مورد اشاره شده است:

✓ شاید این تحقیق اولین پژوهشی باشد که سیستمی جامع و کامل و مستقل طراحی کرده است، از این نظر که این سیستم از گردآوری داده‌ها تا انتشار گزارشات نهایی را دربرمی‌گیرد.

✓ پژوهش‌های مشابه پیشین اغلب از داده‌های آماده استفاده کرده‌اند و معدود پژوهش‌هایی که فرآیند گردآوری داده‌های واقعی را طی کرده‌اند نیز غالبا این داده‌ها را از سمت کاربر استخراج نکرده‌اند؛ لکن به دلیل این‌که در این پژوهش یکی از اهداف پژوهشگران کمک به تصمیم‌گیری در مقوله‌ی فیلترینگ بوده است، نیاز مبرمی به استخراج داده‌های وب‌گردی از سمت کاربر داشته‌اند؛ و بدین جهت بوده است که فرآیند زمان‌بر طراحی و توسعه‌ی افزونه‌ها و داشبورد شخصی وب‌گردی، طراحی و پیاده شده است.

✓ تفکیک محتوایی صفحات وب در پژوهش‌های پیشین اغلب فقط با کمک الگوریتم‌های متن‌کاوی انجام شده بود. منتهی یکی از مشکلات که در زمان توسعه‌ی همچین سیستمی خود را به شکل حاد نشان خواهد داد، سنگین بودن پردازش هنگام ورود اطلاعات به پایگاه‌داده است. این سنگینی اغلب باعث نیاز به سرورهای فوق‌العاده قوی خواهد شد که خود هزینه‌ی زیادی را به مجموعه تحمیل خواهد نمود. و در غیر این صورت موجب از دست رفتن بخشی از داده‌های ورودی خواهد شد. به همین دلیل پژوهشگران در این تحقیق قبل از متن‌کاوی روی صفحات اینترنتی، آن‌ها را از فیلتری از تکنیک‌های حریصانه رد کرده‌اند و با همین کار، پردازش را تا حد بسیار خوبی سبک کرده‌اند.

✓ در پژوهش‌های پیشین اغلب از پارامترهای کم و مشابهی به منظور تفکیک صفحات اینترنتی و مراجعات کاربران به وب‌سایت‌ها استفاده کرده‌بودند، لکن پژوهشگران در این تحقیق پارامترهای متنوعی برای این تفکیک مورد توجه قرار داده‌اند.

✓ پژوهش‌های پیشین اغلب فقط یک خروجی تعریف کرده و در انتها فقط همان را منتشر کرده‌اند. لذا پژوهشگران در این تحقیق به‌منظور نشان دادن کارایی‌های مختلف این سیستم، گزارشاتی کاملا متنوع آماده و منتشر کرده‌اند.

در انتهای این مقاله پیشنهاداتی خطاب به پژوهشگران، سازمان‌ها و همچنین مراجع تصمیم‌گیر بالادستی مطرح می‌کنیم:

- تفکیک محتوایی صفحات اینترنتی فارسی‌زبان نیاز به پژوهشی مجزا و کامل دارد، علاوه بر اعمال حریصانه‌ی پیشنهادی در این پژوهش و الگوریتم‌های دسته‌بندی متون، بسیاری از توابع مورد استفاده در متن‌کاوی –از قبیل تشخیص کلمات غیرمفهومی، تشخیص اجزای کلام، استخراج عبارات اسمی، تحلیل‌گر صرفی، تشخیص مرجع ضمایر و عبارات ارجاعی، استخراج کلمات کلیدی و تشخیص نقل قول– و همچنین توجه به ساختار *html* صفحات وب نیز می‌تواند مورد استفاده قرار گرفته و دقت دسته‌بندی نهایی را بالا ببرد.

- از دیگر پیشنهادات پژوهشگران این تحقیق به دیگر پژوهشگران این حوزه، این است که یک گام به جلو برداشته

و یک سیستم پیشنهاددهنده طراحی و پیاده‌سازی کنند؛ بدین صورت که پس از مدت کوتاهی از نصب افزونه توسط کاربر جدید، با توجه به روند وب‌گردی وی، دسته‌ی وی مشخص شده و طبق آن و با توجه به صفحات جدیدی که توسط دیگر کاربران مورد بازدید قرار می‌گیرد، به این کاربر در هر زمان تعدادی صفحه‌ی اینترنتی به عنوان پیشنهاد ارائه شود. طراحی و پیاده‌سازی این سیستم علاوه بر خروجی تحقیق حاضر، نیاز به تعداد بسیار بالایی کاربر دارد تا بتوان صفحات داغ هر روز در هر دسته و زیردسته را تشخیص دهد تا بتواند در زمان پیشنهاد صفحه‌ی اینترنتی به کاربران مختلف، این مورد نیاز در نظر گرفته شود.

- پیشنهاد می‌شود با پیاده‌سازی این سیستم در ادارات، دانشکده‌ها و مجموعه‌های مختلف علاوه بر رصد لحظه‌به‌لحظه‌ی وب‌گردی اعضای مجموعه، در بازه‌های زمانی مختلف گزارشاتی گرفته شده و نتایج این گزارشات را در تصمیم‌گیری‌های خود دخالت دهید.

- امید است سیستم پیشنهادی را به صورت شکیل و با تبلیغ بهتر محصول ارائه‌شده –داشبورد شخصی وب‌گردی– پیاده‌سازی کرده و به این صورت با وسیع شدن دامنه‌ی کاربران افزونه، اطلاعات خروجی قابل اعتنا و قابل اتکایی دریافت گردد. این اطلاعات مدیران را در بسیاری از تصمیم‌گیری‌های کلان مرتبط با فضای مجازی - از تعرفه‌گذاری تا مبحث فیلترینگ– و حتی فضای واقعی - سبک زندگی، ایجاد راهبردهای فرهنگی و اجتماعی و ... یاری خواهد کرد.

References

Abtahi, A., Elahi, F., & Yousefi-Zenouz, R. (2017). An Intelligent System for Fraud Detection in Coin Futures Market's Transactions of Iran Mercantile Exchange Based on Bayesian Network. Journal Of Information Technology Management, 9(1), 1-20. (Persian)

Ali, W., & Alrabighi, M. (2016). Web Users Clustering Based on Fuzzy C-MEANS. VAWKUM Transactions on Computer Sciences, 11(1), 1-09.

Anitha, A. (2016). An Efficient Agglomerative Clustering Algorithm for Web Navigation Pattern Identification. Circuits and Systems, 7(09), 2349.

Attardi, G., Gullì, A., & Sebastiani, F. (1999). Automatic Web page categorization by link and context analysis. In Proceedings of THAI (Vol. 99, No. 99, pp. 105-119).

Chen, R. C., & Hsieh, C. H. (2006). Web page classification based on a support vector machine using a weighted vote schema. Expert Systems with Applications, 31(2), 427-435.

Cheung, Y. M., & Jia, H. (2013). Categorical-and-numerical-attribute data clustering based on a unified similarity metric without knowing cluster number. Pattern Recognition, 46(8), 2228-2238.

Ciarelli, P. M., Oliveira, E., & Salles, E. O. (2014). Multi-label incremental learning applied to web page categorization. Neural Computing and Applications, 24(6), 1403-1419.

Cooley, R., Mobasher, B., & Srivastava, J. (1999). Data preparation for mining world wide web browsing patterns. Knowledge and information systems, 1(1), 5-32.

Deshmukh, S. M., & Adhiya, K. P. (2016). A Review on Finding Users Navigation Behavior Using Web Mining Algorithm. International Journal of Scientific Research in Science, Engineering and Technology (IJSRSET), 2(6), 708-712.

Dharmarajan, K., & Dorairangaswamy, M. A. (2016). Discovering User Pattern Analysis from Web Log Data using Weblog Expert. Indian Journal of Science and Technology, 9(42).

Dumais, S., & Chen, H. (2000, July). Hierarchical classification of Web content. In Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval (pp. 256-263). ACM.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. AI magazine, 17(3), 37.

Han, J., Pei, J., & Kamber, M. (2011). Data mining: concepts and techniques. Elsevier.

Huang, Z. (1997, February). Clustering large data sets with mixed numeric and categorical values. In Proceedings of the 1st pacific-asia conference on knowledge discovery and data mining,(PAKDD) (pp. 21-34).

Huang, Z. (1998). Extensions to the k-means algorithm for clustering large data sets with categorical values. Data mining and knowledge discovery, 2(3), 283-304.

Kosala, R., & Blockeel, H. (2000). Web mining research: A survey. ACM Sigkdd Explorations Newsletter, 2(1), 1-15.

Kwon, O. W., & Lee, J. H. (2000, November). Web page classification based on k-nearest neighbor approach. In Proceedings of the fifth international workshop on on Information retrieval with Asian languages (pp. 9-15). ACM.

Larose, D. (2014). Discovery Knowledge in Data: An Introduction to Data Mining, 2nd edition. John Wiley-Interscience.

Mladenic, D. (1998). Turning yahoo into an automatic web-page classifier.

Niknam, F., & Niknafs, A. (2016). Improving Text Mining Methods in Market Prediction via Prototype Selection Algorithms. Journal Of Infor-

MCS), 2014 International Conference on (pp. 595-600).

Singh, S., & Aswal, M. S. (2016, October). Towards a framework for web page recommendation system based on semantic web usage mining: A case study. In Next Generation Computing Technologies (NGCT), 2016 2nd International Conference on (pp. 329-334). IEEE.

Wan, M., Jönsson, A., Wang, C., Li, L., & Yang, Y. (2012). Web user clustering and Web prefetching using Random Indexing with weight functions. Knowledge and information systems, 33(1), 89-115.

Xie, X., & Wang, B. (2016). Web page recommendation via twofold clustering: considering user behavior and topic relation. Neural Computing and Applications, 1-9.

Xu, J., & Liu, H. (2010). Web user clustering analysis based on KMeans algorithm. In 2010 International Conference on Information, Networking and Automation (ICINA).

mation Technology Management, 8(2), 415-435. (Persian)

Özel, S. A. (2011). A web page classification system based on a genetic algorithm using tagged-terms as features. Expert Systems with Applications, 38(4), 3407-3415.

Peng, X., & Choi, B. (2002). Automatic web page classification in a dynamic and hierarchical way. In Data Mining, 2002. ICDM 2003. Proceedings. 2002 IEEE International Conference on (pp. 386-393). IEEE.

Raj, A. J., Francis, F. S., & Benadit, P. J. (2016). Optimal Web Page Classification Technique Based on Informative Content Extraction and FA-NBC. Computer Science and Engineering, 6(1), 7-13.

Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. Communications of the ACM, 18(11), 613-620.

Sanoja, A., & Gancarski, S. (2014, April). Block-o-matic: A web page segmentation framework. In Multimedia Computing and Systems (IC-

Human Information Interaction. Vol. 4, No. 2
http://hii.khu.ac.ir
Original Article

Designing a System for Trend Analysis of Users in Websites Surfing in Iran Using Data Mining and Text Mining Algorithms

Babak Sohrabi: Professor of Information Technology Management, Faculty of Management, University of Tehran, Tehran, Iran (Corresponding author). bsohrabi@ut.ac.ir
Iman Raeesi Vanani: Assistant Professor of Industrial Management, Faculty of Management and Accounting, Allameh Tabataba'i University, Tehran, Iran.
Mohammadreza Khorrami: Graduate of Information Technology Management, Faculty of Management, University of Tehran, Tehran, Iran.

Abstract
Background and Aim: As of the entrance of web surfing to the lifestyle of a vast majority of people in the society and the need for a more accurate social and cultural policy making in the field, authors intended to analyze the behavior of the society users in viewing different websites so as to help politicians and practitioners.
Methods: Design science research method is used in this research. The data sample of research consists of all available users that surf Iranian and foreign websites. For gathering data from various active users, some add-ons were designed and published over browsers so as to gather sufficient data.
Results: Through the utilization of text mining algorithms, the browsed webpages were differentiated and using data mining algorithms, the pages were categorized and interpreted.
Conclusion: Finally, a comprehensive system was designed for the analysis of internet users' web browsing trends which contains the data gathering phase and innovative report preparation that can be used as an effective sample for analysis, design, and implementation of web-based analytical systems.

Keywords: Web page categorization, Action clustering, Web surfing, Trend analysison.