
بسم الله الرحمن الرحيم

دانشگاه تهران

پردیس دانشکده‌های فنی

دانشکده مهندسی برق و کامپیوتر

شناسایی صفحات وب هرز فارسی

نگارش

الهه ربانی

استاد راهنما

دکتر آزاده شاکری

پایان‌نامه برای دریافت درجه کارشناسی‌ارشد در رشته

مهندسی کامپیوتر ـ گرایش نرم‌افزار

شهریور ۱۳۹۳

دانشگاه تهران

پردیس دانشکده‌های فنی

دانشکده مهندسی برق و کامپیوتر

پایان نامه برای دریافت درجه کارشناسی ارشد در رشته مهندسی کامپیوتر

عنوان:

شناسایی صفحات وب هرز فارسی

نگارش: الهه ربانی

این پایان‌نامه در تاریخ ۱۳۹۳/۰۶/۱۲ در مقابل هیأت داوران دفاع گردید و مورد تصویب قرار گرفت.

معاون آموزشی و تحصیلات تکمیلی پردیس دانشکده‌های فنی: دکتر علی افضلی کوشا

رئیس دانشکده مهندسی برق و کامپیوتر: دکتر شاهرخ فرهنگی

معاون پژوهشی و تحصیلات تکمیلی: دکتر ناصر معصومی

استاد راهنما: خانم دکتر آزاده شاکری

عضو هیأت داوران: خانم دکتر فتانه تقی‌یاره

عضو هیأت داوران: آقای دکتر بهروز مینایی

تعهدنامه اصالت اثر

اینجانب الهه ربانی تایید می‌کنم که مطالب مندرج در این پایان نامه حاصل کار و پژوهش اینجانب بوده و به دستارودهای پژوهشی دیگران که در این نوشته از آنها استفاده شده است مطابق مقررات ارجاع گردیده است. به‌علاوه این پایان نامه قبلا برای احراز هیچ مدرک هم سطح یا بالاتر ارائه نشده است.

نام و نام خانوادگی دانشجو: الهه ربانی

امضای دانشجو:

تقدیم به

پدر و مادر عزیزم

که از نگاهشان صلابت

از وجودشان عشق

و از صبرشان ایستادگی

را آموختم

تقدیر و تشکر

مراتب تشکر و قدردانی خود را نسبت به تمام کسانی که مرا در انجام این پایان‌نامه یاری کرده‌اند، خصوصاً استاد گرامی و ارجمندم، سرکار خانم دکتر شاکری که رهنمودهای ایشان همواره راهگشای پیچیدگی‌های این پژوهش بوده است، ابراز می‌دارم.

همچنین، از تمامی دوستانم در آزمایشگاه سیستم های هوشمند اطلاعات که با حضور و محبتشان مرا یاری نمودند تشکر و قدردانی می‌نمایم.

چکیده

با توجه به رشد روزافزون اطلاعات موجود در وب، موتورهای جست‌وجو در بازیابی اطلاعات مورد نیاز کاربران از میان حجم زیادی از اطلاعات نقشی اساسی ایفا می‌کنند. با بررسی رفتار کاربر در اینترنت مشاهده شده است که بیشترین بازدید از یک صفحه وب، به واسطه نتایج اولیه بازیابی شده توسط موتورهای جست‌وجو می‌باشد. با توجه به این امر، ایده هرزنویسی در وب با هدف افزایش رتبه صفحات هرز در میان نتایج موتورهای جست‌وجو مطرح شد. برای شناسایی و مقابله با این صفحات روش‌هایی ارائه شده است که می‌توان آن‌ها را به سه دسته کلی روش‌های مبتنی بر محتوا، روش‌های مبتنی بر پیوند و روش‌های مبتنی بر داده‌های جانبی تقسیم نمود. در این پژوهش تمرکز بر روی دو روش اصلی مبتنی بر محتوا و مبتنی بر پیوند و همچنین ترکیب این دو روش به منظور شناسایی وب‌گاه‌های هرز می‌باشد.

از آن‌جایی که عملکرد موتورهای جست‌وجو در شناسایی وب‌گاه‌های هرز فارسی پایین می‌باشد، در این پژوهش پس از ساخت یک مجموعه داده‌ای مناسب شامل وب‌گاه‌های هرز و معتبر فارسی، به بررسی و تحلیل تعدادی از ویژگی‌های محتوایی برای شناسایی وب‌گاه‌های هرز فارسی می‌پردازیم. سپس با ارائه چندین ویژگی محتوایی جدید و استفاده از روش‌های انتخاب ویژگی، کارایی رده‌بندی وب‌گاه‌ها را افزایش می‌دهیم. در ادامه، یک سامانه جدید شناساگر هرز وب فارسی را ارائه می‌دهیم که از مدل بهبود یافته کیف کلمات برای استخراج ویژگی‌ها استفاده می نماید و نسبت به روش‌های محتوایی پیشین کارایی بالاتری دارد. با توجه به گسترش استفاده از الگوریتم‌های مبتنی بر پیوند در روش‌های هرزنویسی، تعدادی از الگوریتم‌های مهم در این زمینه را مورد بررسی قرار داده و دو الگوریتم جدید ارائه می‌دهیم که بسیاری از نقاط ضعف الگوریتم‌های پیشین را ندارند. در الگوریتم اول برای بهبود انتشار امتیاز اعتماد در گراف وب، از سه سیاست انتخاب بهینه گره‌های بذر، وزن‌دهی به یال‌های گراف برای مشخص کردن میزان اعتبار یال‌ها، و بسط دوره‌ای گره‌های بذر استفاده می‌شود. در الگوریتم دوم با استفاده از انتشار امتیاز هرز، هم‌زمان به صورت پیش‌رو و پس‌رو در سراسر گراف وب، کیفیت رتبه‌بندی وب‌گاه‌های هرز را بهبود می‌دهیم. در آخر نیز به منظور بهبود کیفیت رتبه‌بندی وب‌گاه‌ها روشی پیشنهاد داده می‌شود که برای انتشار امتیاز وب‌گاه‌ها، از احتمال اعتبار و هرز بودن محتوایی وب‌گاه‌ها در تمام بخش‌های گراف استفاده می‌نماید.

در پایان این پژوهش، به منظور ارزیابی روش‌ها و بررسی میزان کارایی آن‌ها، آزمایش‌های مربوطه انجام شده است. نتایج آزمایش‌ها نشان می‌دهد که روش‌های ارائه شده در مقایسه با روش‌های قبلی، از کارایی و دقت بالاتری برخوردار هستند.

واژه‌های کلیدی: *هرزنویسی، هرز وب، شناسایی هرز، انتشار برچسب، ویژگی‌های محتوایی.*

فهرست مطالب

فهرست جدول‌ها

فهرست شکل‌ها

فصل ۱

مقدمه

مفهوم هرز[۱] به طور کلی برای هر سامانه[۲] اطلاعاتی، که می‌تواند دارای ارزش محتوایی پایینی باشد به‌کار می‌رود؛ مانند رایانامه[۳]، وب‌نوشت[۴]، وب و نظرهای کاربران در محیط‌های اجتماعی مجازی. در این میان، مفهوم هرز وب[۵] برای اولین بار در سال ۱۹۹۶ مطرح شد [۱]. از آن پس تعاریف متعددی برای هرزنویسی وب[۶] ارائه شد که در یک تعریف جامع می‌توان آن را به هرگونه عملیاتی نسبت داد که در میان نتایج بازیابی شده توسط موتورهای جست‌وجو[۷] باعث افزایش امتیاز و رتبه[۸] صفحات نامرتبط با پرس‌وجو[۹] می‌شود [۲]. همچنین در تعریفی دیگر که توسط Perkins ارائه شده است [۳]، هرزنویسی وب به هرگونه اقدامی گفته می‌شود که با هدف فریب الگوریتم‌های بازیابی اطلاعات[۱۰] و الگوریتم‌های رتبه‌دهی[۱۱] انجام می‌شود. یک صفحه یا وب‌گاه[۱۲] هرز، صفحه یا وب‌گاهی است که دارای محتویات و خصوصیات خاصی می‌باشد که با هدف هرزنویسی ایجاد شده‌اند. این‌گونه از صفحات نه‌تنها اطلاعات مفیدی را در اختیار کاربران قرار

[۱] spam
[۲] system
[۳] email
[۴] weblog
[۵] web spam
[۶] web spamming
[۷] search engines
[۸] rank
[۹] query
[۱۰] Information Retrieval (IR)
[۱۱] ranking
[۱۲] website

نمی‌دهند، بلکه با کاهش کیفیت رتبه‌بندی نتایج پرس‌وجوها، کاربران را از دسترسی به منابع اطلاعاتی مفید و مرتبط با پرس‌وجو محروم می‌نمایند.

با توجه به تاثیر صفحات هرز در کارایی موتورهای جست‌وجو، هرز وب پس از مدت کوتاهی به عنوان یکی از مسائل مهم در صنعت موتورهای جست‌وجو شناخته شد [۴]. همچنین، رشد سریع و روز افزون وب، ساده بودن کار با ابزارهای ایجاد محتوا در وب (مانند ویکی‌پدیا، بسترهای نرم‌افزاری[۱]، انجمن‌های گفتگو[۲] و وب‌نوشت‌ها) و کاهش هزینه‌های نگهداری وب‌گاه‌ها موجب شده است که هرزنویسی در وب به سرعت افزایش پیدا کند و با گذشت زمان از روش‌های جدیدتری برای این منظور استفاده شود که با استفاده از راه‌حل‌های گذشته نتوان صفحات هرز را شناسایی کرد. با توجه به این مهم و همچنین اثرات نامطلوبی که صفحات هرز در کاهش کارایی موتورهای جست‌وجو و در نتیجه کاهش رضایت کاربران دارند، اخیرا تمام شرکت‌های طراح و پشتیبان موتورهای جست‌وجو، شناسایی اطلاعات غیرمفید در وب و مقابله با آن را به عنوان یکی از مسائل مهم در کار خود اعلام کرده‌اند [۵]. این امر موجب شده است پژوهشگرانی که در این زمینه فعالیت می‌کنند به دنبال یافتن روش‌هایی جدید برای حل مشکلات ناشی از هرز وب باشند.

از جمله مشکلاتی که هرز وب ایجاد می‌کند این است که صفحات هرز کیفیت نتایج جست‌وجو را پایین آورده و وب‌گاه‌های قانونی و معتبر را از امتیاز و سودی که در نبود صفحات هرز بدست می‌آورند محروم می‌کند. همچنین وجود این صفحات موجب می‌شود که اعتماد کاربران نسبت به توسعه‌دهنده یک موتور جست‌وجو کاهش یابد و از موتور جست‌وجوی دیگری استفاده نمایند. از طرف دیگر، وجود هرز وب، شرکت‌های توسعه دهنده موتورهای جست‌وجو را وادار می‌کند که هزینه‌های مالی و محاسباتی قابل توجهی را متحمل شوند. به‌علاوه، وب‌گاه‌های هرز با هدف ایجاد و انتشار انواع بدافزار[۳]ها و ویروس‌ها و ترویج محتویات و اطلاعات غیرمفید و نادرست ایجاد می‌شوند که برای کاربران اینترنت تهدیدی جدی محسوب می‌شود. در سال ۲۰۰۵، مجموع ضررهای مالی ناشی از وجود هرز وب حدود ۵۰ میلیارد دلار [۶] و این ضرر در سال ۲۰۰۹ برابر با ۱۳۰ میلیارد دلار [۷] اعلام شده است. طی بررسی‌های جامعی که Ntoulas و

[۱] platforms
[۲] discussion forums
[۳] malware

همکاران در سال ۲۰۰۶ [۸] بر روی پیکره بزرگی از صفحات وب انجام داده‌اند، مقدار هرز وب را برای زبان انگلیسی ۱۳/۸٪، زبان ژاپنی ۹٪، آلمانی ۲۲٪، و فرانسه ۲۵٪ گزارش کرده‌اند. آن‌ها همچنین نشان داده‌اند که ۷۰٪ از صفحات در دامنه[۱] biz. و ۲۰٪ از صفحات در com. صفحات هرز می‌باشند.

از جمله دلایل اصلی پیدایش هرز وب را می‌توان تعداد محدود نتایج بازیابی شده توسط موتورهای جست‌وجو و نمایش نتایج بر اساس رتبه‌بندی دانست و این‌که کاربران معمولا فقط صفحاتی را بررسی می‌نمایند که در اوایل این فهرست نمایش داده می‌شوند. Silverstein و همکاران [۹] طی تحقیقاتی نشان داده‌اند که برای ۸۰ درصد از پرس‌وجوها، کاربران فقط بر روی سه تا پنج نتیجه اول تلیک[۲] می‌کنند. در تحقیقی دیگر، Jansen و Spink [۱۰] نشان داده‌اند که حدود ۸۰ درصد از کاربرانی که از موتورهای جست‌وجو استفاده می‌کنند حداکثر سه صفحه اول نتایج بازیابی شده را بررسی می‌نمایند. بنابراین صفحاتی که بتوانند در فهرست صفحات بازیابی شده توسط موتورهای جست‌وجو رتبه بالاتری را بدست آورند، از این طریق ترافیک بیشتری را به سمت وب‌گاه خود هدایت می‌کنند. برای بسیاری از وب‌گاه‌های تجاری، افزایش تعداد کاربران بازدیدکننده از وب‌گاه، به منزله افزایش تبلیغات، فروش و سایر سودهای تجاری می‌باشد. اگرچه سایت‌های معتبر برای افزایش مشتریان خود از روش‌های بهینه‌سازی موتور جست‌وجو[۳] استفاده می‌کنند، اما در این میان هرزنویسان[۴] با سوء استفاده از این روش‌ها در صدد افزایش رتبه صفحه خود بیشتر از آن‌چه که استحقاقش را دارند هستند. آن‌ها برای رسیدن به این هدف از روش‌های مختلفی از جمله ایجاد تغییراتی در محتوای صفحه، ایجاد پیوند[۵]های جعلی و تغییر مسیر به صفحه‌ای دیگر برای افزایش رتبه صفحات هرز خود استفاده می‌کنند.

برای مقابله با انواع روش‌های هرزنویسی راه‌حل‌هایی ارائه شده است که می‌توان آن‌ها را بسته به نوع اطلاعاتی که استفاده می‌کنند در مجموع به سه دسته کلی روش‌های شناسایی[۶] هرز وب مبتنی بر محتوا، مبتنی بر پیوند و مبتنی بر داده‌های پراکنده تقسیم نمود. در این پژوهش تمرکز ما بر روی دو روش اصلی

[۱] domain
[۲] click
[۳] Search Engine Optimization (SEO)
[۴] spammers
[۵] link
[۶] detection

مبتنی بر محتوا و مبتنی بر پیوند و همچنین ترکیب این دو روش می‌باشد. در ادامه انواع روش‌های مقابله با هرز وب را به طور مختصر معرفی می‌نماییم.

روش‌های مبتنی بر محتوا

روش‌های هرزنویسی مبتنی بر محتوا از جمله روش‌های اولیه و بسیار گسترده در زمینه هرزنویسی وب می‌باشند که اساس عملکرد آن‌ها ایجاد تغییراتی در محتوای صفحات وب است. با توجه به این‌که موتورهای جست‌وجو از مدل‌های مختلف بازیابی اطلاعات مبتنی بر محتوای صفحه مانند مدل فضای برداری[۱] [۱۱]، BM25 [۱۲] و مدل‌های زبانی آماری[۲] [۱۳] برای رتبه‌بندی صفحات استفاده می‌کنند، این روش‌ها با پیدا کردن نقاط ضعف این مدل‌ها موتورهای جست‌وجو را فریب داده و رتبه صفحات هرز را افزایش می‌دهند. بنابراین، در روش‌های مقابله با هرز مبتنی بر محتوا با بررسی ویژگی[۳]های محتوایی صفحات و تحلیل و مقایسه مدل زبانی[۴] آن‌ها میزان احتمال هرز بودن صفحات تخمین زده می‌شود.

بخش زیادی از الگوریتم‌هایی که در این روش‌ها استفاده می‌شوند الگوریتم‌های وابسته به زبان هستند و کارایی آن‌ها برای صفحات به زبان‌های مختلف، متفاوت می‌باشد. با وجود مطالعات زیادی که پیرامون کارایی روش‌های مبتنی بر محتوا بر روی صفحات وب انگلیسی انجام شده است، تاکنون هیچ مطالعه قابل توجهی بر روی شناسایی وب‌گاه‌های هرز فارسی صورت نگرفته است و موتورهای جست‌وجو نیز همچنان در شناسایی این نوع از صفحات هرز عملکرد پایینی دارند. این مهم در حالی است که W^3Techs [۱۴] میزان وب‌گاه‌های فارسی را در سال ۲۰۱۴ حدود ۰/۸ درصد تخمین زده است که این مقدار به سرعت در حال افزایش می‌باشد. در نتیجه، بهبود دقت روش‌های شناسایی وب‌گاه‌های هرز فارسی می‌تواند تاثیر بسزایی در کارایی موتورهای جست‌وجو داشته باشد. همچنین به دلیل نبود پژوهشی قابل توجه در این زمینه، تاکنون هیچ مجموعه داده‌ای[۵] برچسب[۶] خورده استانداردی از وب‌گاه‌های هرز فارسی ایجاد نشده است. بنابراین،

[۱] vector space
[۲] statistical language models
[۳] feature
[۴] language model
[۵] dataset
[۶] label

در این پژوهش ابتدا به جمع‌آوری مجموعه‌ای از وب‌گاه‌های هرز فارسی پرداخته‌ایم و وب‌گاه‌ها را به صورت دستی برچسب‌گذاری کرده‌ایم. توضیحات کامل ساخت این مجموعه داده‌ای در بخش ۱.۱.۳ ارائه شده است. پس از ساخت مجموعه‌ای از وب‌گاه‌های فارسی برچسب خورده، تاثیر انواع ویژگی‌های محتوایی و روش‌های مختلف یادگیری ماشین[۱] بر روی شناسایی وب‌گاه‌های هرز فارسی بررسی شده و سپس تعدادی ویژگی محتوایی جدید برای شناسایی این نوع از وب‌گاه‌ها پیشنهاد شده است. برای بررسی میزان کارایی این ویژگی‌ها در شناسایی وب‌گاه‌های هرز فارسی، نمودار توزیع پراکندگی وب‌گاه‌ها و احتمال هرز بودن آن‌ها با توجه به مقادیر مختلف هر ویژگی به طور مجزا ارائه شده است. در مرحله بعد با استفاده از روش انتخاب ویژگی[۲] χ^2-test و روش حذف پس‌رو[۳]، مجموعه‌ای از ویژگی‌ها به عنوان ویژگی‌های بهینه انتخاب شده‌اند. در نهایت وب‌گاه‌های موجود در مجموعه داده‌ای فارسی، با استفاده از مجموعه ویژگی‌های بهینه و الگوریتم جنگل تصادفی[۴] رده‌بندی[۵] شده‌اند. نتایج بخش ارزیابی نشان می‌دهد که کارایی ویژگی‌های محتوایی معرفی شده در این پژوهش با کارایی تعداد زیادی از ویژگی‌های محتوایی پیشین برابری می‌کند. به علاوه، هزینه محاسباتی ویژگی‌های جدید کمتر از بسیاری از ویژگی‌های محتوایی پیشین می‌باشد. همچنین بررسی نتایج حاصل از رده‌بندی وب‌گاه‌ها پس از انتخاب مجموعه ویژگی‌های بهینه نشان می‌دهد که استفاده از تعداد ویژگی‌های بیشتر، نه‌تنها سبب بهبود نتایج رده‌بندی نمی‌شود، بلکه ترکیب برخی از این ویژگی‌های محتوایی با یکدیگر در مواردی کارایی رده‌بندی را کاهش می‌دهد. از طرف دیگر بررسی میزان تاثیر هر یک از ویژگی‌های محتوایی به تنهایی در کارایی رده‌بندی نشان می‌دهد که خصوصیات وب‌گاه‌های هرز به گونه‌ای می‌باشد که با استفاده از یک ویژگی محتوایی به تنهایی نمی‌توان تمام آن‌ها را شناسایی کرد.

با توجه به این‌که با استفاده از مجموعه ویژگی‌های بهینه، همچنان امکان شناسایی تعدادی از وب‌گاه‌های هرز وجود ندارد، در ادامه با بررسی محتوای وب‌گاه‌های هرز و معتبر فارسی، یک سامانه جدید برای شناسایی هرز وب فارسی پیشنهاد شده است. در این سامانه از یک روش مبتنی بر مدل جدیدی به نام کیف کلمات

[۱] machine learning
[۲] feature selection
[۳] backward elimination
[۴] random forest
[۵] classification

هرز[۱] (BOSW) که نسخه‌ای تغییر یافته از مدل ساده کیف کلمات[۲] می‌باشد، برای انتخاب ویژگی‌ها استفاده شده است. نتایج بخش ارزیابی نشان می‌دهد که این روش در تشخیص هرز وب فارسی از لحاظ دو معیار فراخوانی[۳] و دقت[۴] نسبت به روش‌های محتوایی پیشین بهتر عمل می‌کند.

روش‌های مبتنی بر پیوند

روش‌های هرزنویسی مبتنی بر پیوند، روش‌هایی هستند که در ساختار گرافی بین صفحات وب تغییراتی را ایجاد می‌کنند. این تغییرات به نحوی می‌باشند که در نهایت منجر به افزایش امتیازی می‌شوند که با استفاده از الگوریتم‌های رتبه‌بندی مبتنی بر گراف مانند PageRank [۱۵]، HITS [۱۶] و TrustRank [۱۷] به صفحات داده می‌شود. یکی از این تغییرات ایجاد دهکده پیوندی[۵] [۱۸، ۱۹] می‌باشد. یک دهکده پیوندی مجموعه‌ای از صفحات وب است که با پیوندهای زیادی به یکدیگر ارتباط دارند و می‌توانند با فریب الگوریتم‌های رتبه‌دهی امتیاز همه یا تعدادی از صفحات درون آن دهکده پیوندی را افزایش دهند. در سال ۲۰۰۰، Davision [۲۰] پس از ارائه تعریفی جامع برای هرزنویسی وب به روش مبتنی بر پیوند، روش‌هایی برای شناسایی پیوندهای جعلی در وب پیشنهاد داده است. طبق تعریف Davision، هرزنویسی مبتنی بر پیوند به روش‌هایی گفته می‌شود که سعی در ایجاد پیوندهای جعلی بین صفحات دارند تا بدین طریق بتوانند رتبه پیوندی صفحات هرز را افزایش دهند. در این روش‌ها ساختار گرافی بین صفحات به نحوی تغییر داده می‌شود که با اعمال الگوریتم‌های رتبه‌بندی، صفحات هرز بتوانند رتبه‌ای بالاتر از آن چه که استحقاقش را دارند بدست آورند.

با توجه به وجود این نوع از روش‌های هرزنویسی، به تدریج الگوریتم‌هایی برای مقابله با این نوع از صفحات هرز پیشنهاد داده شدند که اساس عملکرد آن‌ها بر روی ساختار کلی گراف و ارتباطات پیوندی بین صفحات است و از روش‌هایی برای انتشار امتیاز اعتماد[۶] یا عدم اعتماد[۷] صفحات، پیدا کردن دهکده‌های

[۱] bag-of-spam-words
[۲] bag-of-words
[۳] recall
[۴] precision
[۵] link farm
[۶] trust
[۷] distrust

پیوندی، حذف پیوندهای مشکوک و تشخیص رفتارهای غیرعادی در الگوریتم‌های رتبه‌بندی مبتنی بر پیوند استفاده می‌کنند.

در این پژوهش، دو الگوریتم مبتنی بر پیوند برای رده‌بندی وب‌گاه‌ها پیشنهاد شده است که اساس کار آن‌ها بر روش انتشار برچسب می‌باشد. در الگوریتم WorthyRank با استفاده از روش‌هایی مانند انتخاب بهینه وب‌گاه‌های بذر[۱]، بسط دوره‌ای مجموعه وب‌گاه‌های بذر و همچنین وزن‌دهی به یال[۲]های گراف وب و حذف یال‌های جعلی، دقت رتبه‌بندی صفحات وب بهبود داده می‌شود. در ادامه نیز الگوریتمی به نام JunkyRank معرفی شده است که احتمال هرز بودن وب‌گاه‌ها را هم‌زمان به دو صورت پیش‌رو[۳] و پس‌رو[۴] در گراف وب انتشار می‌دهد. نتایج نشان می‌دهد که این الگوریتم در بازیابی و رتبه‌بندی صفحات هرز به نسبت روش‌های قبلی از دقت بالاتری برخوردار است. برای آزمایش روش‌های ارائه شده از دو مجموعه داده‌ای استاندارد WebSpamChallengeII-CorpusI و WEBSPAM-UK2007 استفاده شده است. از دو معیار ضریب هرز[۵] و ضریب اطمینان[۶] نیز برای ارزیابی این روش‌ها استفاده شده است.

پس از معرفی و تحلیل هر یک از روش‌های مبتنی بر محتوا و مبتنی بر پیوند به طور مجزا، در آخر برای بهبود دقت و کارایی رتبه‌بندی وب‌گاه‌ها، روشی جدید ارائه شده است که از ترکیب روش محتوایی با روش پیوندی استفاده می‌کند. در این روش ابتدا تعدادی از وب‌گاه‌ها به همراه مجموعه وب‌گاه‌های بذر گراف، به عنوان داده‌های آموزش انتخاب شده و برچسب‌گذاری می‌شوند. سپس با استفاده از ویژگی‌های محتوایی استخراج شده از این مجموعه وب‌گاه‌ها، یک رده‌بند[۷] محتوایی ساخته می‌شود. در مرحله بعد با استفاده از این رده‌بند برای هر یک از وب‌گاه‌های گراف، احتمال اعتبار و هرز بودن محتوایی محاسبه می‌شود. در نهایت با استفاده از الگوریتم‌های پیوندی ارائه شده در این پژوهش و با در نظر گرفتن احتمال هرز و اعتبار محتوایی هر گره[۸]، میزان اعتبار وب‌گاه‌های بذر در کل گراف انتشار داده می‌شود. با استفاده از این روش،

---

[۱] seed
[۲] edge
[۳] forward
[۴] backward
[۵] spam factor
[۶] confidence factor
[۷] classifier
[۸] node

برای هر وب‌گاه یک امتیاز نهایی اعتبار و یا عدم اعتبار محاسبه شده و وب‌گاه‌ها بر اساس این امتیاز نهایی رتبه‌بندی می‌شوند. نتایج بخش ارزیابی نشان می‌دهد که این روش نسبت به سایر روش‌های ارائه شده در این پژوهش از دقت بالاتری برخوردار است.

روش‌های مبتنی بر داده‌های پراکنده

این دسته از روش‌ها برای شناسایی صفحات هرز از اطلاعاتی استفاده می‌کنند که فراتر از محتویات ایستا[۱] ی صفحات و یا ساختار گرافی آن‌ها می‌باشد. یکی از این روش‌ها بررسی و تحلیل رفتار کاربران در مرورگرهای اینترنتی می‌باشد. برای مثال سابقه جست‌وجوهای کاربران در محیط وب و بررسی تاریخچه تلیک بر روی پیوندهای مختلف می‌تواند اطلاعات خوبی را درباره میزان مفید بودن وب‌گاه‌های مختلف و ارتباط آن‌ها با یکدیگر در اختیار قرار دهد. تحلیل اطلاعات مربوط به HTTP (در سمت سرویس‌گیرنده[۲] و در سمت سرویس‌دهنده[۳])، محاسبه مدت زمان استفاده کاربران از صفحات هر وب‌گاه و سایر اطلاعاتی که به طور برخط[۴] توسط موتورهای جست‌وجو جمع‌آوری و به‌روز می‌شود، روش‌های خوبی برای تخمین میزان اعتبار وب‌گاه‌های مختلف هستند. با توجه به این‌که موتورهای جست‌وجو امکان دسترسی به این اطلاعات را برای کاربران عادی فراهم نمی‌کنند، در این پژوهش از بررسی این دسته از روش‌ها صرف‌نظر شده است.

در ادامه این نوشتار، در فصل ۲، تعدادی از کارهای پیشین مرتبط با انواع روش‌های شناسایی و مقابله با هرز وب را مرور می‌نماییم. شرح کامل انواع روش‌های پیشنهادی برای رده‌بندی وب‌گاه‌ها در فصل ۳ ارائه می‌شود. در این فصل ابتدا انواع ویژگی‌های محتوایی را که برای رده‌بندی وب‌گاه‌های فارسی استفاده کرده‌ایم توضیح داده و تعدادی ویژگی محتوایی جدید را معرفی می‌نماییم. همچنین روش محتوایی دیگری را معرفی می‌کنیم که از مدل کیف کلمات هرز برای استخراج ویژگی‌های محتوایی استفاده می‌کند. در ادامه این فصل، دو الگوریتم مبتنی بر پیوند و نیمه‌سرپرست[۵] را معرفی می‌کنیم که به نسبت روش‌های قبلی کارایی و دقت بالاتری دارند. در نهایت نیز الگوریتمی را پیشنهاد می‌دهیم که از ترکیب روش محتوایی با روش‌های

[۱] static
[۲] client-side
[۳] server-side
[۴] online
[۵] semi-supervised

پیوندی برای رتبه‌بندی وب‌گاه‌ها استفاده می‌کند. در فصل ۴ ابتدا توضیحاتی را درباره مجموعه داده‌ای‌های استفاده شده در این پژوهش ارائه می‌دهیم، سپس پس از معرفی معیارهای ارزیابی، تمام روش‌های معرفی شده در این پژوهش را مورد آزمایش و ارزیابی قرار می‌دهیم. همچنین تحلیل‌های مربوط به نتایج هر آزمایش نیز در این فصل بیان شده است. در نهایت در فصل ۵، ضمن جمع‌بندی مطالب و نتیجه‌گیری، پیشنهادهایی برای کارهای آینده در این زمینه ارائه می‌دهیم.

فصل ۲

پژوهش‌های پیشین

در این فصل، روش‌هایی که تاکنون برای هرزنویسی وب و همچنین تشخیص صفحات هرز معرفی شده‌اند را در چهار گروه اصلی دسته‌بندی و بررسی می‌کنیم. بدین منظور، ابتدا به مطالعه روش‌های مبتنی بر محتوا می‌پردازیم، سپس مطالعات انجام شده بر روی روش‌های مبتنی بر پیوند را بررسی می‌نماییم. در بخش بعد، کارهایی را مرور می‌کنیم که از اطلاعات مربوط به رفتار کاربران در محیط وب برای شناسایی صفحات هرز استفاده می‌کنند. در آخر نیز تعدادی از پژوهش‌های اخیر را که از ترکیب روش‌های قبلی استفاده می‌کنند، مورد مطالعه قرار می‌دهیم.

۱.۲ روش‌های مبتنی بر محتوا

از جمله فعالیت‌هایی که در رابطه با تشخیص صفحات وب هرز بر مبنای محتوای صفحات صورت گرفته است، تحقیقات انجام شده توسط Fetterly و همکاران [۵، ۲۱–۲۳] است. آن‌ها در پژوهشی [۲۱] در سال ۲۰۰۴، نشان داده‌اند که استفاده از تحلیل‌های آماری در شناسایی انواع صفحات وب هرز به خصوص صفحاتی که به طور خودکار توسط ماشین ایجاد شده‌اند بسیار کاربرد دارد. بدین منظور، تعدادی ویژگی برای شناسایی صفحات هرز معرفی کرده‌اند که عبارتند از: تعداد پیوندهای ورودی به هر صفحه، تعداد

پیوندهای خروجی، تعداد کلمات غیرنشانه‌گذاری[۱] هر صفحه، میزان شباهت محتوای صفحات که با استفاده از الگوریتم Shingling [۲۴] محاسبه شده و خوشه‌بندی[۲] آن‌ها، ویژگی‌های خاص مربوط به آدرس اینترنتی صفحات، بررسی آدرس IP مربوط به هر نام میزبان[۳]، و سرعت و درصد تغییر محتوای صفحات در طول زمان [۲۲]. سپس با انجام آزمایش‌هایی نشان داده‌اند که به ازای هر یک از ویژگی‌های تعریف شده، داده‌های با مقادیر برون‌هشته[۴] با احتمال زیادی نمایانگر صفحات هرز می‌باشند. نتایج آزمایش بر روی آدرس اینترنتی صفحات نشان می‌دهد که آدرس صفحات هرز معمولا دارای تعداد زیادی کلمه، نقطه، خط تیره و عدد می‌باشد. همچنین با بررسی آدرس IP و نام میزبان صفحات نشان داده‌اند که چندین وب‌گاه هرز با نام میزبان متفاوت، دارای یک آدرس IP مشترک هستند. خصوصیت دیگری که در این مقاله به عنوان یکی از ویژگی‌های صفحات هرز بیان شده است، سرعت زیاد و درصد بالای تغییرات محتوای ایستای صفحات می‌باشد.

در پژوهشی دیگر، Fetterly و همکاران [۲۳] روشی را معرفی کرده‌اند که می‌تواند دسته‌ای از صفحات هرز را شناسایی کند که دارای محتوای تکراری در سطح عبارت هستند. محتوای این صفحات با کنار هم گذاشتن بخش‌هایی از محتوای صفحات دیگر ساخته می‌شود. روشی که در این پژوهش برای شناسایی این دسته از صفحات استفاده شده است، از بسیاری از ویژگی‌های روش انگشت‌نگاری[۵] Robin [۲۵، ۲۶] بهره می‌گیرد. در سال ۲۰۰۶، در راستای تکمیل کارهای قبلی [۲۱، ۲۳]، Ntoulas و همکاران [۸] تعدادی ویژگی محتوایی را برای شناسایی خودکار صفحات هرز معرفی کرده و میزان کارایی هر یک را به طور مجزا در تشخیص این نوع از صفحات بررسی کرده‌اند. سپس، با ترکیب ویژگی‌ها و استفاده از روش‌های مختلف یادگیری ماشین، دقت رده‌بندی را افزایش داده‌اند. نتایج آن‌ها نشان می‌دهد که بالاترین میزان دقت رده‌بندی برای الگوریتم درخت تصمیم[۶] است که توانسته‌اند با استفاده از دو روش bagging [۲۷] و boosting [۲۸] دقت این الگوریتم را نیز افزایش دهند. در هر دو روش مجموعه‌ای از رده‌بندها ایجاد و در نهایت با یکدیگر ترکیب می‌شوند تا یک رده‌بند با دقت بالاتر ایجاد کنند. هفت سال

[۱] non-markup
[۲] clustering
[۳] host name
[۴] outlier
[۵] fingerprinting
[۶] decision tree

بعد، Prieto و همکاران [۲۹] یک سامانه شناسایی و تحلیل هرز به نام SAAD معرفی کرده‌اند که بر مبنای تعدادی ویژگی محتوایی کار می‌کند. در این سامانه علاوه بر استخراج ویژگی‌های محتوایی معرفی شده در مقاله [۸]، تعدادی ویژگی محتوایی جدید نیز معرفی و استخراج شده است. این ویژگی‌ها عبارتند از: متوسط طول لغات (با حذف ایست‌واژه[۱]ها و ابربرچسب[۲]ها)، درصد کد پویا[۳] و ایستای موجود در صفحه، تعداد عکس‌های بدون توضیح و برچسب، تعداد ایست‌واژه‌ها و تعدادی ویژگی دیگر. در این مقاله نیز پس از تحلیل کارایی هر یک از ویژگی‌های ارائه شده به طور مجزا، با ترکیب این ویژگی‌ها و ویژگی‌های معرفی شده در مقاله [۸] و استفاده از الگوریتم درخت تصمیم به همراه دو روش bagging و boosting به دقت بالاتری رسیده‌اند.

Sydow و همکاران [۳۰]، کاربرد انواع روش‌های یادگیری ماشین را در شناسایی صفحات وب هرز بررسی کرده‌اند. بدین منظور آن‌ها از تعدادی ویژگی زبانی استفاده کرده و میزان کارایی آن‌ها را در شناسایی صفحات هرز مورد بررسی قرار داده‌اند. در یک پژوهش کامل‌تر، Piskorski و همکاران [۳۱] برای بهبود شناسایی صفحات هرز تعدادی ویژگی زبانی جدید پیشنهاد داده‌اند. بسیاری از این ویژگی‌های زبانی که شامل تعداد کلمات متمایز موجود در متن صفحه، تعداد اسم‌ها و فعل‌ها، تعداد جملات مجهول، میزان گوناگونی[۴] در لغات و محتوای متن و میزان نقل قول‌های مستقیم و غیرمستقیم می‌باشند با استفاده از ابزارهای پردازش زبان طبیعی[۵] استخراج شده‌اند. Martinez-Romo و همکاران [۳۲]، با بهره‌گیری از مدل زبانی صفحات و محاسبه میزان اختلاف آن‌ها، صفحات وب هرز را شناسایی می‌کنند. این روش قبلا در کاربردهای بسیاری مانند شناسایی و مقابله با وب‌نوشت‌های هرز [۳۳] استفاده شده است. با بهره‌گیری از این روش، در مقاله [۳۲]، مدل زبانی‌متن پیوند[۶] و کلمات موجود در آدرس الکترونیکی صفحه مبدا با مدل زبانی عنوان و بدنه صفحه مقصد مقایسه شده و اختلاف Kullback-Leibler (KL) Divergence بین این دو مدل زبانی محاسبه می‌شود. در صورتی که میزان این اختلاف از یک آستانه مشخص بیشتر باشد صفحه

[۱] stopword
[۲] Meta tag
[۳] dynamic
[۴] diversity
[۵] Natural Language Processing (NLP)
[۶] anchor text

مقصد با احتمال بالایی هرز است. از جمله مزیت‌های این روش این است که نیازی به داده‌های آموزش[۱] نداشته و هزینه و زمان اجرای کمتری دارد. برخلاف بسیاری از کارهای قبلی که از محتویات غیرنشانه‌گذاری صفحات برای تشخیص هرز بودن یا نبودن آن‌ها استفاده کرده‌اند، Urvoy و همکاران [۳۴]، از ویژگی‌های مبتنی بر ساختار صفحات HTML برای رده‌بندی صفحات وب استفاده می‌کنند. در این روش‌ها، طی یک مرحله پیش‌پردازش تمامی محتویات قابل مشاهده صفحات حذف شده و قالب اصلی صفحه HTML نگه داشته می‌شود. سپس برای پیدا کردن گروه‌هایی از صفحات که از لحاظ ساختار صفحه مشابه یکدیگر می‌باشند از روش انگشت‌نگاری [۲۵، ۲۶] صفحات و سپس خوشه‌بندی آن‌ها استفاده می‌کنند.

در دسته دیگری از کارها [۳۵–۳۸]، از روش مدل موضوعی[۲] برای شناسایی صفحات وب هرز استفاده می‌شود. مدل موضوعی یک مدل آماری است که مشخص می‌کند یک سند یا مجموعه‌ای از اسناد در رابطه با چه موضوع یا موضوعاتی می‌باشد. یکی از ساده‌ترین روش‌های ایجاد مدل موضوعی، روش Latent Dirichlet Allocation (LDA) [۳۹] می‌باشد که با استفاده از آن می‌توان برای هر سند مجموعه‌ای از موضوعات مرتبط با آن سند به همراه احتمال مربوط بودن هر یک به آن سند را مشخص کرد. Biro و همکاران [۳۵]، از LDA برای شناسایی هرز وب استفاده می‌کنند. آن‌ها برای هر وب‌گاه، یک سند به صورت کیف کلمات ایجاد کرده و الگوریتم LDA را بر روی هر دسته از صفحات هرز و معتبر اجرا می‌کنند. با این روش، در مرحله آموزش مجموعه‌ای از موضوع‌های هرز و معتبر ایجاد می‌شود. در مرحله آزمون، برای هر صفحه مدل موضوعی آن را ایجاد کرده و در صورتی که احتمال موضوعی هرز آن از یک میزان مشخصی بالاتر باشد آن صفحه به عنوان صفحه هرز شناسایی می‌شود. در پژوهشی دیگر [۳۶]، از توزیع احتمال موضوعی در کنار تحلیل گوناگونی در سطح حروف، کلمات و جملات برای شناسایی هرز وب استفاده می‌شود. با استفاده از این روش، صفحاتی که دارای مدل موضوعی با توزیع احتمال یکنواخت هستند به احتمال زیاد صفحات معتبر می‌باشند و به میزانی که این توزیع از یکنواخت بودن فاصله می‌گیرد احتمال هرز بودن صفحات افزایش می‌یابد. در سال ۲۰۱۲ نیز Dong و Zhou [۳۷]، با معرفی تعدادی معیار گوناگونی موضوعی، پژوهشی مشابه [۳۶] را انجام داده‌اند. Suhara و همکاران [۳۸]، با بهره‌گیری از روشی که

[۱] train data
[۲] topic model

در [۴۰] ارائه شده است، به‌جای ایجاد مدل موضوعی از روی کل یک سند، هر جمله را یک سند مجزا در نظر گرفته و برای آن یک مدل موضوعی ایجاد می‌کنند. تفاوتی که کار آن‌ها با کار قبلی دارد این است که به‌جای استفاده از توزیع احتمال موضوعی، هر جمله را با موضوعی که بالاترین احتمال را دارد جایگزین می‌کنند. بدین ترتیب، هر سند را به دنباله‌ای از موضوعات تبدیل کرده و سپس میزان گوناگونی موضوعی داخل هر صفحه را بررسی می‌کنند. سپس با انجام آزمایش‌هایی نشان می‌دهند که میزان بی‌نظمی موضوعی در صفحات هرز بیشتر است.

در سال‌های اخیر، مطالعات زیادی در زمینه شناسایی صفحات وب هرز عربی انجام شده است. بدین منظور، Wahsheh و همکاران [۴۱] مجموعه‌ای از صفحات وب عربی را تهیه کرده و به صورت دستی برچسب‌گذاری کرده‌اند. در پژوهش‌های دیگر [۴۲، ۴۳] کارایی برخی از ویژگی‌های محتوایی بر روی این مجموعه صفحات وب عربی، بررسی شده است. سپس با استفاده از الگوریتم‌های یادگیری ماشین، نشان داده‌اند که درخت تصمیم بالاترین میزان دقت را در شناسایی هرز وب عربی دارد. در سال ۲۰۱۲، Al-Kabi و همکاران [۴۴]، مجموعه‌ای کامل‌تر از صفحات وب هرز عربی را جمع‌آوری و برچسب‌گذاری کرده‌اند. سپس با استفاده از روش‌های محتوایی و الگوریتم‌های یادگیری ماشین صفحات را رده‌بندی کرده‌اند. Al-Kabi و همکاران [۴۵] در آخرین پژوهش خود در سال ۲۰۱۴، با معرفی یک سامانه برخط شناسایی هرز وب عربی که از هر دو ویژگی محتوایی و پیوندی و همچنین از اطلاعات مربوط به بازخوردهای کاربران استفاده می‌کند، توانسته‌اند دقت تشخیص هرز وب عربی را افزایش دهند.

۲.۲ روش‌های مبتنی بر پیوند

با توجه به مطالعاتی که در راستای انجام این پژوهش، بر روی روش‌های مبتنی بر پیوند صورت گرفته است، این گروه از روش‌ها را می‌توان به سه دسته اصلی تقسیم کرد. دسته اول پژوهش‌هایی [۴۶–۴۸] هستند که ابتدا تمام ویژگی‌های پیوندی صفحات را استخراج کرده، سپس از الگوریتم‌های یادگیری ماشین برای رده‌بندی صفحات استفاده می‌کنند. دومین دسته از این روش‌ها، الگوریتم‌هایی [۴۹–۵۱] هستند

که از ایده بهینه‌سازی برچسب‌ها بر اساس ساختار گرافی وب و همچنین روش‌های graph regularization [۵۲–۵۴] برای تشخیص صفحات هرز استفاده می‌کنند. الگوریتم‌های دسته سوم [۱۵، ۵۵–۵۹]، ارتباطات ساختاری (مانند فاصله صفحات در گراف وب، ارجاع‌های مشترک، شباهت) بین صفحات برچسب‌دار و سایر صفحات در گراف وب را بررسی کرده و از میزان و نحوه ارتباط آن‌ها با یکدیگر برای رده‌بندی صفحات استفاده می‌کنند. در تعدادی از این روش‌ها [۱۷، ۶۰–۶۳]، با داشتن برچسب تعدادی از صفحات بذر و همچنین روش انتشار برچسب، میزان هرز بودن یا نبودن سایر صفحات محاسبه می‌شود. با توجه به این‌که تمرکز اصلی ما در این پژوهش بر روی این دسته از روش‌های مبتنی بر پیوند می‌باشد، در ادامه توضیحات بیشتری را پیرامون این الگوریتم‌ها ارائه می‌دهیم.

ایده اصلی در این دسته از الگوریتم‌های نیمه‌سرپرست این است که با داشتن برچسب مجموعه‌ای از گره‌های گراف (صفحات وب) و با استفاده از قوانین انتشار مختلف، سعی می‌کنند تا برچسب سایر صفحات را پیش‌بینی کنند. یکی از کارهای اولیه در این زمینه الگوریتم TrustRank [۱۷] است که بر اساس این فرضیه ارائه شده است که صفحات معتبر معمولا به صفحات معتبر ارجاع می‌دهند. در این روش، مجموعه‌ای از صفحات معتبر به عنوان گره‌های بذر انتخاب شده و امتیاز اعتماد از طریق پیوندهای خروجی و با استفاده از الگوریتم PageRank شخصی‌سازی شده[۱] از این مجموعه صفحات معتبر به سایر مجموعه صفحات گراف انتشار داده می‌شود. Gyongyi و همکاران [۱۷] نشان داده‌اند که در تشخیص صفات وب هرز، الگوریتم TrustRank نسبت به الگوریتم PageRank دقت و کارایی بیشتری دارد. در سال ۲۰۰۶، Wu و همکاران [۶۴] الگوریتم TrustRank موضوعی[۲] را پیشنهاد داده‌اند که صفحات بذر را بر اساس موضوعشان دسته‌بندی می‌کنند. سپس الگوریتم TrustRank برای هر موضوع به طور جداگانه اجرا می‌شود. در نهایت با ترکیب این امتیازها، میزان اعتبار نهایی صفحات مشخص می‌شود. در پژوهشی دیگر [۵۶] برای بهبود الگوریتم TrustRank، روشی پیشنهاد شده است که کیفیت پیوندهای یک صفحه را مستقل از کیفیت خود صفحه در نظر گرفته است و هدف آن تخصیص امتیاز پیوندی به هر صفحه بر اساس کیفیت پیوندهای آن می‌باشد. همچنین Chen و همکاران [۶۵] با بررسی پیوندهای متغیر و میزان

[۱] personalized
[۲] topical

تغییرات در ساختار پیوندی گراف بین صفحات، توانسته‌اند الگوریتم TrustRank را بهبود دهند. روش دیگر، الگوریتم Anti-TrustRank [۶۰] است که امتیاز هرز بودن را از مجموعه‌ای از صفحات هرز اولیه به سایر صفحات انتشار می‌دهد. نحوه انتشار امتیاز هرز در این الگوریتم مانند الگوریتم TrustRank است. با این تفاوت که برای اجرای این الگوریتم، امتیاز هرز بودن مجموعه صفحات بذر، در خلاف جهت یال‌های گراف به سایر صفحات انتشار داده می‌شود. این امر به این دلیل است که اساس این الگوریتم بر اساس این فرضیه است که صفحاتی که به صفحات هرز ارجاع می‌دهند با احتمال زیادی هرز هستند. Krishnan و همکاران [۶۰] نشان داده‌اند که الگوریتم Anti-TrustRank نسبت به الگوریتم TrustRank دقت بالاتری دارد و توانایی آن نیز در تشخیص صفحات هرزی که PageRank بالاتری دارند بیشتر است.

Wu و همکاران [۴۹] روشی را ارائه داده‌اند که می‌تواند صفحات هرز درون دهکده‌های پیوندی را تشخیص دهد. در این روش ابتدا برای انتخاب مجموعه صفحات بذر، به ازای هر صفحه، تعداد صفحات مشترک بین صفحاتی که به آن صفحه ارجاع داده‌اند و صفحاتی که توسط آن صفحه ارجاع داده شده‌اند محاسبه می‌شود. سپس در صورتی که این تعداد، از یک آستانه مشخص بیشتر باشد، آن صفحه به عنوان صفحه بذر انتخاب می‌شود. این روش انتخاب بذر بر اساس ارتباط زیاد صفحات درون دهکده پیوندی تعریف شده است. سپس با توجه به این امر که اگر صفحه هرزی به تعداد زیادی صفحه هرز پیوند داده باشد، با احتمال زیادی خود یک صفحه هرز است، در هر مرحله صفحاتی که تعداد ارجاعاتشان به صفحات بذر بیشتر از آستانه تعریف شده باشد، به مجموعه صفحات بذر اضافه می‌شوند. این الگوریتم تا جایی ادامه پیدا می‌کند که از آن پس، صفحه جدیدی به عنوان صفحه هرز شناسایی نشود.

پس از معرفی روش‌های نیمه‌سرپرست انتشار برچسب که به مجموعه‌ای از صفحات برچسب خورده به عنوان صفحات بذر الگوریتم نیاز دارد، پژوهش‌هایی در رابطه با چگونگی انتخاب بذر اولیه و تاثیراتی که بر کارایی نهایی الگوریتم دارد انجام شد. Jiang و همکاران [۶۶] نشان داده‌اند که در صورتی که اندازه بذر در الگوریتم‌های انتشار برچسب مانند الگوریتم TrustRank کم باشد، نتیجه نهایی به سمت صفحات بذر سوگیری[۱] می‌کند. کم بودن تعداد صفحات بذر اولیه باعث می‌شود که درصد بیشتری از امتیاز صفحات،

[۱] bias

به صفحات بذر اختصاص یابد و در نتیجه در نتایج نهایی با احتمال زیادی این صفحات در رتبه‌های بالاتر قرار بگیرند. آن‌ها پیشنهاد داده‌اند که تعداد صفحات بذر باید با توجه به این‌که چند نتیجه اول برای کاربر مهم است، انتخاب شود. بدین صورت که در شرایطی که کاربر تعداد نتایج بیشتری را بررسی می‌کند، تعداد صفحات بذر اولیه نیز باید بیشتر باشد. از طرف دیگر افزایش اندازه بذر اولیه، هزینه زمانی را افزایش می‌دهد. با توجه به این مهم، Zhang و همکاران [۶۷] یک روش خودکار برای افزایش تعداد صفحات بذر اولیه پیشنهاد داده‌اند. در این الگوریتم که به صورت دوره‌ای تکرار می‌شود، با داشتن مجموعه صفحات معتبر اولیه، در هر تکرار، تمام صفحاتی که به هر صفحه پیوند داده‌اند بررسی شده و در صورتی که در این میان تعداد صفحات معتبر از مقدار آستانه بیشتر باشد، صفحه مبدا به عنوان صفحه معتبر شناخته شده و به بذر اولیه اضافه می‌شود. شرط بیشتر بودن از آستانه به این دلیل است که با پیدایش وب‌نوشت‌ها و انجمن[۱] های مختلف که امکان نوشتن نظرات و مطالب گوناگون را به کاربرهای مختلف می‌دهند، هرزنویسان با قرار دادن پیوند صفحه هرز خود در این بخش از صفحات معتبر، سعی در فریب روش‌های پیوندی دارند. بدین ترتیب، اگرچه ممکن است صفحات هرز با استفاده از چنین روش‌هایی [۶۸] تعدادی پیوند از صفحات معتبر به صفحات خود ایجاد کنند، اما به دلایلی از جمله مشکل هزینه و زمان‌بر بودن این کار، امکان ایجاد چنین پیوندهایی به تعداد زیاد وجود ندارد. در [۶۴] نیز Wu و همکاران، مرحله انتخاب بذر را به عنوان مهم‌ترین بخش الگوریتم TrustRank دانسته‌اند و معتقدند که چگونگی انتخاب بذر اولیه در کارایی نهایی الگوریتم تاثیر می‌گذارد. آن‌ها با در نظر گرفتن این امر که صفحات وب در رابطه با موضوعات مختلفی هستند، نشان داده‌اند که نتایج نهایی روش TrustRank به سمت صفحاتی سوگیری می‌شود که موضوعشان با موضوع صفحات بذر یکسان است. برای جلوگیری از این مشکل، صفحات بذر را از موضوعات مختلف انتخاب کرده و آن‌ها را بر اساس موضوعشان به گروه‌های مختلف تقسیم می‌کنند. سپس الگوریتم انتشار برچسب را برای هر گروه با موضوع مجزا اجرا کرده و امتیاز نهایی صفحات را از ترکیب امتیازهایی که برای هر موضوع بدست آورده‌اند، محاسبه می‌کنند. Zhao و همکاران [۶۹] نشان داده‌اند که در انتخاب بذر اولیه، استفاده از صفحات هرز با تعداد پیوندهای ورودی بیشتر، کارایی الگوریتم Anti-TrustRank را افزایش می‌دهد. در سال ۲۰۰۹، Zhang و همکاران [۷۰، ۷۱] برای محاسبه اعتبار صفحات، الگوریتمی به

[۱] forum

نام CPV معرفی کرده‌اند که اساس روش آن مبتنی بر الگوریتم HITS [۱۶] می‌باشد و از پیوندهای دوجهته[1] برای بهبود امتیازدهی به صفحات استفاده می‌کند. در این الگوریتم نیز مانند HITS در هر تکرار الگوریتم، دو امتیاز مبتنی بر hub و authority به ترتیب به نام‌های HVRank و AVRank برای هر صفحه محاسبه شده و در نهایت از ترکیب این دو امتیاز برای محاسبه میزان اعتبار صفحات استفاده می‌شود. آن‌ها همچنین نشان داده‌اند که استفاده از پیوندهای دوجهته، مشکل سوگیری نتیجه به سمت صفحات بذر را حل می‌کند.

تعدادی از پژوهش‌ها برای افزایش کارایی روش‌های شناسایی صفحات وب هرز، از انتشار هم‌زمان امتیاز اعتماد و عدم اعتماد در گراف وب استفاده کرده‌اند. در [۶۱]، پس از اجرای هر یک از الگوریتم‌های TrustRank و Anti-TrustRank به صورت مجزا بر روی گراف، نتایج بدست آمده به صورت خطی با یکدیگر ترکیب شده و صفحات بر اساس آن رتبه‌بندی می‌شوند. Zhang و همکاران [۷۲] الگوریتمی به نام TDR را پیشنهاد داده‌اند که برای هر صفحه دو امتیاز TRank و Drank را که مشخص کننده میزان اعتبار و عدم اعتبار صفحات است، محاسبه می‌کند. اساس کار این الگوریتم مانند الگوریتم‌های TrustRank و Anti-TrustRank است، با این تفاوت که در هر مرحله، برای انتشار امتیاز اعتماد (عدم اعتماد) از صفحه مبدا به صفحه مقصد، امتیاز صفحه مبدا به میزان احتمال اعتبار (عدم اعتبار) صفحه مقصد انتشار داده می‌شود. پس از آن، در پژوهشی دیگر [۷۳]، روشی به نام GBR معرفی شده است که مشابه الگوریتم TDR می‌باشد، با این تفاوت که برای انتشار امتیاز، به جای بررسی صفحه مقصد، میزان احتمال هرز بودن یا معتبر بودن صفحه مبدا را در نظر می‌گیرد. نویسندگان این مقاله نشان داده‌اند که این روش در مقایسه با روش TDR از کارایی بالاتری برخوردار است.

۳.۲ روش‌های مبتنی بر داده‌های پراکنده

در این دسته از روش‌ها از ویژگی‌های غیر رایج برای شناسایی صفحات وب هرز استفاده می‌کنند. می‌توان این روش‌ها را با توجه به نوع اطلاعاتی که استفاده می‌کنند به چند زیرگروه تقسیم نمود. با توجه به این‌که این دسته از روش‌ها در محدوده تحقیقاتی این پژوهش نمی‌باشند، به توضیح مختصری از تحقیقات مهم

[1] bidirectional

در این زمینه اکتفا می‌کنیم.

تعدادی از پژوهش‌ها [۷۴–۷۶]، برای شناسایی صفحات هرز، از رفتار کاربران در محیط وب و تاریخچه جست‌وجوی صفحات، توسط کاربران استفاده می‌کنند. در [۷۴]، رفتار کاربر در مرورگرها به صورت یک گراف، مدل سازی شده است. در این گراف، یال‌ها نشان دهنده رفتن کاربر از یک صفحه به صفحه دیگر از طریق پیوند درون صفحه مبدا می‌باشند. همچنین زمان توقف کاربر در هر صفحه و احتمال پرش تصادفی از یک صفحه به صفحات دیگر نیز محاسبه شده و در نهایت گراف، به یک فرآیند مارکوف با زمان پیوسته تبدیل می‌شود. روش ارائه شده در [۷۵] نیز مشابه روش [۷۴] می‌باشد، با این تفاوت که گراف مدل‌سازی شده در [۷۵]، اجتماعی از گراف استاندارد وب و گراف ساخته شده از روی رفتار کاربران است. Liu و همکاران [۷۶]، با بررسی‌هایی که بر روی رفتار کاربران در محیط وب انجام داده‌اند، دو ویژگی را برای صفحات هرز تعریف کرده‌اند. خصوصیت اول این است که با توجه به این‌که صفحات هرز دارای محتوای مناسبی نیستند، با فریب موتورهای جست‌وجو سعی دارند که ترافیک بیشتری را به سمت خود جذب کنند. از طرف دیگر، کاربران در صورت مشاهده صفحه هرز، بلافاصله به صفحه دیگری می‌روند. بنابراین برای شناسایی صفحات هرز سه ویژگی جدید معرفی کرده‌اند که عبارتند از: تعداد بازدیدکنندگان از یک صفحه به واسطه موتورهای جست‌وجو، تعداد تلیک بر روی صفحات و زمان بازدید از هر صفحه وب.

در دسته دیگری از پژوهش‌ها [۷۷–۷۹]، از روش‌های بدون‌سرپرست[۱] برای شناسایی هرز وب استفاده می‌شود. الگوریتم‌هایی که در این دسته از روش‌ها استفاده می‌شوند الگوریتم‌های در سمت مشتری برخط هستند که به داده‌های آموزش نیازی ندارند. یکی از کارهای اصلی در این زمینه، [۸۰] است که از مدلی به نام دهکده صفحه[۲] استفاده می‌کند. در این الگوریتم در هر مرحله، به ازای هر صفحه از بین k نزدیک‌ترین همسایه به طور حریصانه[۳]، آن‌هایی که بیشترین تاثیر را در امتیاز PageRank آن صفحه دارند انتخاب کرده و میزان هرز بودن پیوندی را با استفاده از میزان مشارکت مشاهده شده در امتیاز PageRank تقسیم بر حالت بهینه این مقدار محاسبه می‌کنند. در صورتی که برای تمام همسایه‌های یک صفحه، مقدار هرز بودن پیوندی از یک آستانه مشخص بیشتر باشد، آن صفحه به عنوان صفحه هرز در نظر گرفته می‌شود.

[۱] unsupervised
[۲] page farm
[۳] greedy

آن‌ها همچنین، مشابه این الگوریتم را برای محتوای صفحه اجرا می‌کنند. بدین صورت که مقدار هرز بودن محتوایی را، نسبت امتیاز TF-IDF [۸۱] محتوای مشاهده شده از صفحه بر امتیاز TF-IDF بهینه‌ای که از یک صفحه با همان تعداد کلمه بدست می‌آید تعریف می‌کنند.

مجموعه‌ای دیگر از روش‌ها، از اطلاعات نشست[۱] HTTP برای شناسایی هرز وب استفاده می‌کنند. برخی از آن‌ها مانند [۸۲]، از اطلاعات محدودی (مانند مدل‌سازی اطلاعات درون عنوان درخواست‌های در سمت مشتری و در سمت خدمتگزار) استفاده می‌کنند که معمولا نیازی به یادگیری ندارند و بنابراین دقت پایینی دارند. در مقابل، تعداد دیگری از روش‌ها، از اطلاعات بی‌درنگ[۲] نیز برای افزایش دقت یادگیری بهره می‌گیرند. در [۸۳]، روش‌های افزایش تعداد کاربران بازدیدکننده از صفحات هرز به سه نوع مختلف تقسیم‌بندی شده و نمودار توزیع هر یک از انواع آن در وب بررسی شده است. در نهایت، روشی برای شناسایی تغییر مسیر جاوا اسکریپت[۳] که یکی از سخت‌ترین روش‌ها می‌باشد، پیشنهاد داده شده است. در [۸۴] نیز استفاده از ویژگی‌های رتبه‌ـزمان در کنار ویژگی‌های مستقل از پرس‌وجو پیشنهاد شده است. در این مقاله Svore و همکاران از ۳۴۴ ویژگی رتبه‌ـزمان استفاده کرده‌اند که تعدادی از آن‌ها عبارتند از: تعداد کلمات درون پرس‌وجو که در عنوان صفحه استفاده شده است، میزان تکرار کلمات پرس‌وجو در صفحه، تعداد صفحاتی که کلمه پرس‌وجو را درون خود دارند و میزان هم‌پوشانی nـ گرام‌های پرس‌وجو با هر صفحه. نتایج آزمایش‌ها نشان داده است که با اضافه کردن این ویژگی‌های رتبه‌ـزمان می‌توان بدون کاهش مقدار فراخوانی، میزان دقت را تا ۲۵ درصد افزایش داد.

در این میان پژوهش‌هایی نیز بر روی شناسایی هرز تلیک[۴] انجام شده است. هدف از هرز تلیک، ایجاد خطا در اطلاعاتی است که توسط موتورهای جست‌وجو پیرامون پرس‌وجوها و نتایج انتخاب شده از میان آن‌ها، جمع‌آوری می‌شود. با استفاده از این روش، هرزنویسان می‌توانند در داده‌هایی که برای ایجاد توابع رتبه‌بندی استفاده می‌شود خطا ایجاد کنند. بیشتر پژوهش‌های انجام شده در راستای مقابله با این روش‌ها، به دنبال مقاوم کردن الگوریتم‌های یادگیری در برابر این دسته از خطاها می‌باشند. یکی از کارهای

[۱] session
[۲] real-time
[۳] javaScript
[۴] click spam

مهمی که در این زمینه انجام شده است، مقاله [۸۵] توسط Radlinski و همکاران می‌باشد. آن‌ها نشان داده‌اند که توابع رتبه‌بندی شخصی‌سازی شده به دلیل مقاومتشان در برابر خطاها، در شناسایی هرز تلیک بسیار خوب عمل می‌کنند. در [۸۶] نیز، میزان مقاومت تابع رتبه‌بندی بر اساس تلیک، در مقابل خطاهای هرز تلیک بررسی شده است. در پژوهش دیگری [۸۷] نیز، یک مدل رتبه‌بندی ارائه شده است که برای ساخت آن از اطلاعات کاربران و بازخوردهای آن‌ها استفاده می‌شود. در این روش به کاربر این امکان داده می‌شود که خطاهای درون سامانه را شناسایی و گزارش کند.

۴.۲ روش‌های ترکیبی

این دسته از روش‌ها علاوه بر تحلیل ویژگی‌های محتوایی صفحات، ساختار پیوندی صفحات را نیز مورد بررسی قرار می‌دهند. در تعدادی از پژوهش‌هایی [۸۸—۹۱] که در این زمینه انجام شده است، ابتدا تمام ویژگی‌های محتوایی و پیوندی صفحات استخراج شده و سپس از ترکیب آن‌ها برای رده‌بندی صفحات وب استفاده می‌شود. Geng و همکاران [۸۸] در سال ۲۰۰۷، با داشتن مجموعه‌ای از ویژگی‌های محتوایی و پیوندی صفحات موجود در مجموعه داده‌ای WEBSPAM-UK2006، داده‌های آموزش را به صورت تصادفی به چند گروه تقسیم کرده و با استفاده از هر گروه، رده‌بندی را ایجاد کرده‌اند که به ازای هر صفحه آزمون، احتمال هرز بودن آن را پیش‌بینی می‌کند. سپس از ترکیب خروجی این رده‌بندها، برای رده‌بندی نهایی صفحات استفاده می‌کنند. در پژوهش دیگری [۸۹]، با داشتن مجموعه‌ای از ویژگی‌های محتوایی و پیوندی از مجموعه داده‌ای WEBSPAM-UK2007، و با استفاده از روش‌های مختلف انتخاب ویژگی مانند Information Gain (IG)، Chi-Squared (χ^2-test) [۹۲] و CFS[۱] [۹۳]، میزان کارایی هر دسته از ویژگی‌ها و توانایی آن‌ها در شناسایی صفحات هرز بررسی شده است. همچنین در سال ۲۰۱۰، Geng و همکاران [۹۰]، ویژگی‌های چند مقیاسی را معرفی کرده‌اند که از چهار بخش مختلف از اطلاعات مربوط به صفحات وب استخراج شده‌اند و عبارتند از: ویژگی‌های آماری محتوای صفحات، ویژگی‌های مبتنی بر ساختار پیوندی بین صفحات، ویژگی‌های مبتنی بر ساختار پیوندی بین میزبان‌ها و همچنین ویژگی‌های

[۱] Correlation based Feature Selection

TF-IDF کلمات درون صفحات. پس از استخراج این ویژگی‌ها، آن‌ها را با هم ترکیب کرده و با استفاده از روش bagging و درخت تصمیم C4.5 به عنوان رده‌بند پایه، صفحات را رده‌بندی کرده‌اند. در دسته‌ای دیگر از پژوهش‌ها [۹۴–۹۶]، تاثیر روش‌های انتخاب ویژگی و الگوریتم‌های یادگیری ماشین، بر روی دقت شناسایی هرز وب بررسی شده است.

Araujo و Martinez-Romo [۹۷]، یک سامانه شناسایی هرز را معرفی کرده‌اند که از هر دو ویژگی محتوایی و پیوندی به منظور رده‌بندی صفحات وب استفاده می‌کند. بدین منظور، آن‌ها تعدادی ویژگی محتوایی و پیوندی جدید معرفی کرده‌اند. یکی از ویژگی‌هایی که در این پژوهش به منظور شناسایی صفحات هرز بررسی می‌شود، میزان توانایی موتورهای جست‌وجو در بازیابی صفحاتی است که توسط صفحه یا صفحات دیگری به آن‌ها ارجاع داده شده است. در این روش، تعدادی از کلمات درون متن پیوند مربوط به پیوند صفحه ارجاع داده شده، که در صفحه مبدا وجود دارد، به عنوان پرس‌وجو به یک موتور جست‌وجو داده شده و ده نتیجه اول آن بررسی می‌شود. اگر صفحه ارجاع داده شده در فهرست ده صفحه اول بازیابی شده باشد، آن پیوند به عنوان یک پیوند معتبر در نظر گرفته می‌شود. با استفاده از این روش، تعداد پیوندهای معتبر و نامعتبر درون هر صفحه را مشخص کرده و از اختلاف آن‌ها برای شناسایی صفحات هرز استفاده می‌کنند. نمودار مربوط به این ویژگی نشان می‌دهد که نسبت پیوندهای معتبر به نامعتبر در صفحات هرز کمتر از صفحات معتبر می‌باشد. با در نظر گرفتن این امر که صفحات هرز به صفحات معتبر ارجاع می‌دهند اما صفحات معتبر به صفحات هرز ارجاع نمی‌دهند، ویژگی دیگری را نیز برای شناسایی هرز وب معرفی کرده‌اند. برای محاسبه این ویژگی، برای هر صفحه، اختلاف تعداد پیوندهای ورودی از هر یک از صفحات هرز و معتبر و همچنین تعداد پیوندهای خروجی به هر یک از این صفحات را محاسبه می‌کنند. بررسی میزان اختلاف تعداد پیوندهای داخلی با تعداد پیوندهای خارجی صفحات وب نیز نشان داده است که صفحات هرز دارای پیوندهای خارجی بیشتری نسبت به پیوندهای داخلی می‌باشند. ویژگی‌های دیگری که در این مقاله معرفی و بررسی شده است، تعداد پیوندهای منقضی شده یک صفحه و همچنین تعداد پیوندهای یک صفحه که متن پیوند ندارند می‌باشد. در این پژوهش علاوه بر این ویژگی‌ها، تعدادی ویژگی دیگر نیز معرفی شده است که از مدل زبانی بخش‌های مختلف صفحات استفاده می‌کند. در این روش، مدل زبانی متن

پیوند مربوط به صفحه ارجاع داده شده در صفحه مبدا با مدل زبانی عنوان صفحه مقصد مقایسه می‌شود. بدین منظور میزان KL-Divergence بین این دو مدل زبانی محاسبه شده و در صورتی که مقدار آن از یک آستانه مشخص بیشتر باشد، آن صفحه به عنوان صفحه هرز شناسایی می‌شود. همچنین، علاوه بر مدل زبانی متن پیوند صفحه مبدا و عنوان صفحه مقصد، مدل زبانی آدرس، عنوان صفحه، بدنه اصلی، متن پیوند و کلمات اطراف متن پیوند، و همچنین کلمات موجود در ابربرچسب‌ها در هر دو صفحه مبدا و مقصد نیز محاسبه شده و میزان KL-Divergence گروه‌های مختلف آن‌ها با یکدیگر محاسبه می‌شود. Abernety و همکاران [۹۸]، یک الگوریتم یادگیری به نام WITCH معرفی کرده‌اند که در مرحله یادگیری، علاوه بر استفاده از ویژگی‌های محتوایی صفحات، به طور هم‌زمان از ساختار پیوندی آن‌ها نیز استفاده می‌کند. در روش ارائه شده، برای یادگیری رده‌بند خطی، از یک تابع هدف مشابه الگوریتم SVM استفاده شده است.

همان‌طور که بررسی مقاله‌های مرتبط با روش‌های مقابله با هرز وب نشان می‌دهد، هیچ‌گونه پژوهش قابل توجهی بر روی شناسایی هرز وب فارسی انجام نشده است. بنابراین در این پژوهش ابتدا به بررسی کارایی روش‌های مبتنی بر محتوا بر روی وب‌گاه‌های فارسی پرداخته و سپس روش جدیدی را برای شناسایی این نوع از وب‌گاه‌ها معرفی می‌نماییم. در ادامه نیز دو روش جدید برای شناسایی هرز وب مبتنی بر پیوند پیشنهاد داده و نشان می‌دهیم که این دو روش نسبت به سایر روش‌های پیوندی موجود بهتر عمل می‌کنند. در آخر نیز برای بهبود قدرت تشخیص هرز وب، یک روش ترکیبی معرفی می‌نماییم که در مقایسه با سایر روش‌های ترکیبی کارایی بالاتری دارد.

فصل ۳

روش‌های پیشنهادی برای شناسایی هرز وب

پس از بررسی اجمالی روش‌های پیشین در زمینه شناسایی هرز وب، در این فصل، به معرفی تعدادی از روش‌های مبتنی بر محتوا و مبتنی بر پیوند برای شناسایی وب‌گاه‌های هرز پرداخته می‌شود. بدین منظور، ابتدا در بخش ۱.۳، تعدادی از روش‌های شناسایی هرز وب مبتنی بر محتوا را بر روی مجموعه وب‌گاه‌های فارسی بررسی کرده و با معرفی ویژگی‌های جدید و همچنین یک سامانه شناساگر هرز وب فارسی، سعی در ارائه الگوریتمی با کارایی بهتر در رده‌بندی وب‌گاه‌های فارسی داریم. پس از آن، در بخش ۲.۳، برای رده‌بندی وب‌گاه‌ها، دو روش مبتنی بر پیوند را، که بر اساس الگوریتم‌های انتشار برچسب ایجاد شده‌اند، معرفی می‌نماییم. در نهایت در بخش ۳.۳، یک روش ترکیبی جدید پیشنهاد می‌دهیم که برای رده‌بندی وب‌گاه‌ها از هر دو دسته ویژگی محتوایی و پیوندی استفاده می‌کند.

۱.۳ شناساگرهای محتوایی هرز وب فارسی

در این بخش، ابتدا به توضیح مختصری درباره چگونگی ساخت مجموعه داده‌ای از وب‌گاه‌های هرز و معتبر فارسی پرداخته، سپس میزان تاثیر انواع ویژگی‌های محتوایی را در رده‌بندی وب‌گاه‌های فارسی بررسی می‌کنیم. در ادامه نیز تعدادی ویژگی محتوایی جدید برای شناسایی هرز وب فارسی معرفی کرده و کارایی

هر یک را در شناسایی هرز وب فارسی بررسی می‌نماییم. پس از این مرحله، به منظور ایجاد رده‌بندی با کارایی بالاتر و هزینه کمتر، از روش‌های انتخاب ویژگی برای انتخاب موثرترین ویژگی‌ها و همچنین بررسی انواع الگوریتم‌های یادگیری ماشین برای انتخاب مناسب‌ترین رده‌بند استفاده شده است. در نهایت در بخش ۳.۱.۳، سامانه‌ای را ارائه می‌دهیم که از مدل جدیدی به نام BOSW، که در این پژوهش معرفی می‌شود، برای شناسایی وب‌گاه‌های هرز استفاده می‌کند. نتایج نشان می‌دهد که این روش در مقایسه با روش اول از دقت و فراخوانی بالاتری برخوردار است.

۱.۱.۳ ساخت پیکره‌ای از مجموعه وب‌گاه‌های هرز و معتبر فارسی

با توجه به نبود پژوهشی مناسب در زمینه شناسایی هرز وب فارسی و عدم دسترسی به مجموعه داده‌ای استاندارد از وب‌گاه‌های فارسی، در این پژوهش، ابتدا به ساخت مجموعه داده‌ای PersianWebSpam-2013 شامل وب‌گاه‌های فارسی برچسب‌خورده، اقدام نمودیم. در ادامه به شرح مراحل ساخت این مجموعه داده‌ای می‌پردازیم.

انتخاب سطح برچسب‌زنی

برچسب زنی صفحات وب به یکی از دو صورت در سطح میزبان[۱] یا در سطح صفحه[۲] انجام می‌شود. در حالت اول، تمام صفحات مربوط به یک وب‌گاه مشترک برچسب یکسان می‌گیرند، در صورتی که در حالت دوم، هر صفحه مستقل از این‌که مربوط به چه وب‌گاهی است به تنهایی بررسی شده و برچسب‌گذاری می‌شود. اگرچه دقت برچسب زنی در سطح صفحه بیشتر است، اما میزان پوشش‌دهی وب‌گاه‌ها و دامنه‌ها در روش اول بالاتر می‌باشد. بنابراین با توجه به اهمیت پوشش‌دهی وب‌گاه‌های مختلف در این پژوهش، برچسب زنی در سطح میزبان انجام شده است.

[۱] host-level
[۲] page-level

ارائه تعاریف مشخص برای وب‌گاه‌های هرز

امروزه با گسترش انواع مختلف وب‌گاه‌های هرز و معتبر، مرز بین این دو نوع وب‌گاه بسیار کم‌رنگ شده است و امکان ارائه یک تعریف جزئی و دقیق برای وب‌گاه‌های هرز وجود ندارد. با این حال می‌توان این نوع از وب‌گاه‌ها را با توجه به مهم‌ترین خصوصیتی که نمایانگر هرز بودن آن‌ها است، به انواع مختلفی تقسیم کرد و برای هر نوع، یک تعریف نسبتا مشخص ارائه داد. بدین منظور، ابتدا لازم است مطالعاتی در زمینه آشنایی با الگوریتم‌های امتیازدهی به صفحات وب که امروزه توسط موتورهای جست‌وجو استفاده می‌شود انجام گیرد. پس از آن، آشنایی با انواع روش‌های هرزنویسی که برای فریب این الگوریتم‌ها استفاده می‌شوند، می‌تواند به پژوهشگران کمک کند تا با انواع خصوصیات صفحات هرز آشنا شوند. یکی از مطالعات مهمی که بدین منظور در این پژوهش انجام شد، مطالعه راهنمای گوگل[۱] می‌باشد که در آن علاوه بر موارد مطرح شده، انواع وب‌گاه‌های هرز به همراه مثالی از هر کدام ارائه شده‌اند. با الهام گرفتن از این راهنما و همچنین راهنمایی که در http://chato.cl/webspam/datasets/uk2006/guidelines/ توسط Castillo و همکاران [۹۹] در اختیار پژوهشگران قرار گرفته است، و همچنین بررسی خصوصیات تعدادی از وب‌گاه‌های هرز و معتبر فارسی که به صورت تصادفی انتخاب شدند، وب‌گاه‌های هرز فارسی را به چند دسته متفاوت تقسیم نمودیم. شرح انواع وب‌گاه‌های هرز به صورت زیر است:

- صفحاتی که دارای انواع مختلفی از کلیدواژه‌ها[۲] درون متن اصلی، متن پیوند و یا ابربرچسب‌های درون صفحه هستند. در این‌گونه از صفحات، که به اصطلاح از روش انباشتگی کلیدواژه‌ها[۳] برای افزایش رتبه خود استفاده می‌کنند، کلیدواژه‌های مختلفی به دفعات زیاد درون صفحه تکرار شده‌اند. هرچند ممکن است بخش‌هایی از این صفحات، شامل محتوای مفید نیز باشد، اما با توجه به این‌که بخش زیادی از رتبه خود را از طریق روش‌های هرزنویسی بدست می‌آورند، به عنوان صفحات هرز محسوب می‌شوند. شکل ۱.۳ نمونه‌ای از این نوع از صفحات هرز را نشان می‌دهد.
- صفحاتی که دارای تعداد زیادی عکس و تبلیغ انواع محصولات مربوط به کالاهای مختلف و تقلبی

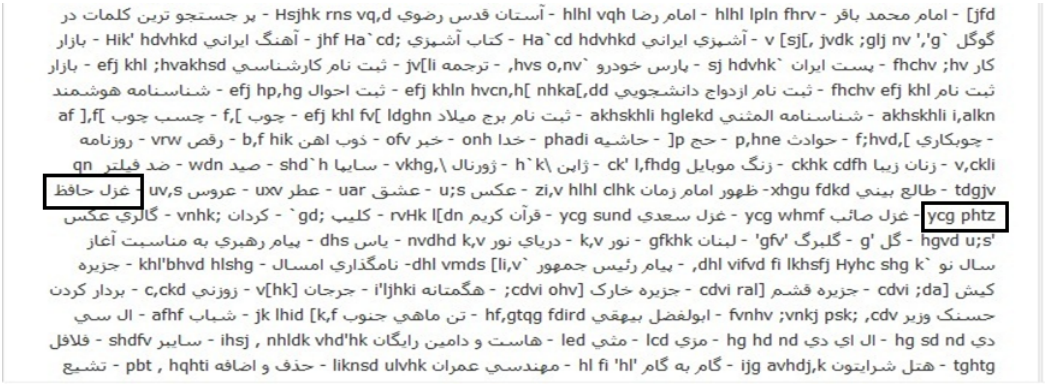
[۱] Google General Guidlines, version 3.27, June 2012.
[۲] keywords
[۳] keyword stuffing

شکل ۱.۳: نمونه‌ای از یک صفحه هرز فارسی که از روش انباشتگی کلیدواژه‌ها برای افزایش رتبه خود استفاده کرده است.

هستند. در برخی از این صفحات، نسبت تعداد ارجاعات به صفحات تبلیغاتی مختلف در آن‌ها، به میزان متن مفید صفحه بسیار زیاد است. نکته‌ای که در مورد این صفحات باید در نظر گرفت این است که صفحات معتبر تبلیغاتی که مربوط به شرکتهای تبلیغاتی معتبر هستند و یا اطلاعات مفیدی هم‌چون مقایسه قیمت و کیفیت کالا، تلفن و آدرس شرکت‌ها و یا سازمان‌های خصوصی و دولتی را در اختیار کاربران قرار می‌دهند، جزء صفحات هرز محسوب نمی‌شوند.

- صفحاتی که شامل کلمات متنوعی هستند که به صورت خودکار و تصادفی توسط ماشین ایجاد شده‌اند. برخی از این کلمات دارای خطای املایی و نوشتاری هستند. برخی از آن‌ها نیز به همراه مدل نوشتاری انگلیسی صفحه کلید ظاهر می‌شوند. با استفاده از این روش، این صفحات می‌توانند در شرایطی که زبان صفحه کلید به اشتباه بر روی حالت انگلیسی است و کاربر کلمات موردنظر را به فارسی وارد می‌کند، رتبه خوبی را بدست آورند. برای مثال، همان‌طور که در شکل ۲.۳ مشاهده می‌کنیم، مدل نوشتاری انگلیسی عبارت «غزل حافظ» به صورت «ycg phtz» نوشته می‌شود.

- صفحاتی که دارای بخش‌های متنی نامرتبط با یکدیگر هستند. یک هرزنویس برای ایجاد چنین صفحه هرزی معمولا بخش‌های مختلفی از متن وب‌گاه‌های معتبر را در صفحه خود رونوشت می‌کند. بنابراین با خواندن متن صفحه می‌توان مشاهده نمود که این صفحه دارای محتوای پیوسته و مفید نمی‌باشد و جملات متوالی از نظر مفهومی ارتباطی با یکدیگر ندارند. همچنین برخی از جملات در این صفحه



شکل ۲.۳: بخشی از یک صفحه هرز که دارای کلیدواژه‌های زیاد به همراه مدل نوشتاری انگلیسی آن‌ها می‌باشد.

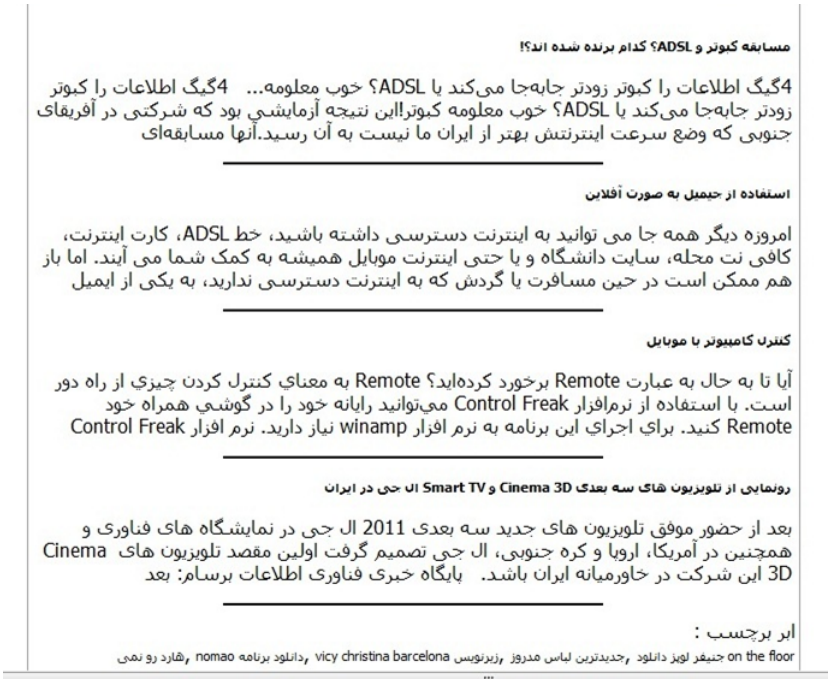
نیمه‌کامل می‌باشند. نمونه‌ای از این صفحات در شکل ۳.۳ ارائه شده است.

- صفحاتی که دارای تعداد زیادی پیوند غیرمفید هستند. این پیوندها معمولا به صفحات نامربوط، صفحات مسدود شده، صفحات هرز و یا صفحات درون همان وب‌گاه که دارای محتوای تکراری هستند، اشاره می‌کنند. در برخی مواقع نیز با تلیک بر روی یک پیوند، کاربر مجددا به همان صفحه هدایت می‌شود. برخی از این پیوندها نیز شبه‌پیوند هستند و در صورت تلیک کردن بر روی آن‌ها اتفاق خاصی نمی‌افتد.

- صفحاتی که دارای محتوای مخفی، مانند بخش‌هایی از متن صفحه و یا پیوندهای پنهان هستند. با استفاده از روش‌هایی مانند استفاده از رنگ پیش‌زمینه صفحه و یا نوشتن متن با اندازه بسیار کوچک، این بخش‌ها از دید کاربر پنهان می‌مانند، اما همچنان توسط موتورهای جست‌وجو نمایه‌سازی[1] شده و امتیاز آن‌ها در محاسبه رتبه صفحه در میان سایر نتایج پرس‌وجو در نظر گرفته می‌شود. برای پیدا کردن این بخش‌ها می‌توان از Ctrl+A استفاده کرد. لازم به ذکر است که وجود اطلاعات پنهان، مانند تاریخ به‌روزرسانی وب‌گاه، که مربوط به مشخصات و یا تنظیمات صفحه است، نشان‌گر یک صفحه هرز نمی‌باشد.

- صفحاتی که URL آنها شامل کلیدواژه‌های زیاد و علامت‌های نگارشی مختلف مانند نقطه و خط تیره است. با توجه به این‌که موتورهای جست‌وجو، آدرس صفحات را نیز نمایه‌سازی می‌کنند و

[1] indexing



شکل ۳.۳: بخشی از یک صفحه هرز که دارای جمله‌های نیمه‌کاره و مطالب نامرتبط با یکدیگر می‌باشد.

از کلمات درون آن‌ها نیز برای رتبه‌دهی به صفحه استفاده می‌نمایند، در این نوع از صفحات با استفاده از کلیدواژه‌ها در آدرس صفحه، سعی می‌کنند رتبه خود را افزایش دهند. برای مثال آدرس http://www.mihanmobile.net/خرید_شال_زنانه_طرح_قلب_و_طاووس.html که مربوط به یک وب‌گاه هرز است و دارای تعدادی کلمه کلیدی که با استفاده از خط تیره جدا شده‌اند، می‌باشد.

- صفحاتی که دارای نظرها و انجمن‌های گفتگوی رونوشت شده می‌باشند. این دسته از صفحات، به ظاهر دارای بخش‌ها و انجمن‌هایی هستند که در آن امکان تبادل نظر و ارسال پیام وجود دارد، اما در اصل چنین امکانی برای کاربران فراهم نشده است. همچنین برخی از این صفحات دارای کلیدهایی به نام «ادامه مطلب»، «خرید کالا»، «دانلود» و «ارسال نظر» هستند که در صورت تلیک کردن بر روی آن‌ها هیچ اتفاقی رخ نمی‌دهد.

- صفحاتی که قبل از بارگذاری، کاربر را به صفحه‌ای با دامنه متفاوت هدایت می‌کنند. همچنین صفحاتی که با تلیک کردن بر روی پیوند آن‌ها، علاوه بر آن صفحه، یک یا چند صفحه تبلیغاتی نیز به طور خودکار باز می‌شود. برای نمونه، با وارد کردن آدرس «http://100cd.ir» در نوار آدرس، مرورگر

به طور خودکار به آدرس «http://shop.sarzamindownload.com» هدایت می شود.

با توجه به این‌که در این پژوهش برچسب‌زنی را در سطح میزبان در نظر گرفته‌ایم، تعریف یک وب‌گاه هرز را به صورتی تقریبا متفاوت در نظر می‌گیریم. مطابق این تعریف، در صورتی که علاوه بر صفحه اصلی یک وب‌گاه، حداقل پنج صفحه از میان ده صفحه‌ای که از آن وب‌گاه به صورت تصادفی انتخاب شده است هرز باشد، آن وب‌گاه به عنوان یک وب‌گاه هرز شناسایی می‌شود. همچنین با بررسی وب‌گاه‌ها، درمی‌یابیم که در صفحات یک وب‌گاه هرز معمولا از ترکیب تعدادی از روش‌های بالا استفاده می‌شود.

جمع‌آوری فهرستی از مجموعه وب‌گاه‌های فارسی

با توجه به این‌که وجود وب‌گاه‌های هرز، نارضایتی کاربران زیادی را به همراه دارد، با جست‌وجو در میان وب‌گاه‌ها و وب‌نوشت‌های مختلف، می‌توان فهرستی از صفحاتی را که توسط کاربران وب به عنوان صفحات هرز گزارش شده‌اند، جمع‌آوری نمود. در این پژوهش علاوه بر استفاده از این روش، به صورت مستقیم نیز فهرستی از وب‌گاه‌های هرز و معتبر را از مراکز تحقیقاتی مختلف جمع‌آوری نمودیم.

همچنین، برای تهیه فهرستی از وب‌گاه‌های معتبر، می‌توان از آدرس مربوط به وب‌گاه‌های دولتی، شرکت‌ها و یا سازمان‌های معتبر استفاده کرد. در این پژوهش از وب‌گاه http://www.i-link.ir که شامل فهرستی از وب‌گاه‌های معتبر در موضوعات مختلف می‌باشد، استفاده نمودیم. در این وب‌گاه، آدرس بسیاری از صفحات معتبر و مفید فارسی، به صورت دسته‌بندی شده قرار دارد. برای مثال وب‌گاه‌های مربوط به دانشگاه‌ها، مراکز آموزشی، مراکز درمانی، بانک‌ها، مراکز تجاری، وب‌گاه‌های خبری و ورزشی، هر یک به صورت مجزا دسته‌بندی شده‌اند. برای جلوگیری از سوگیری وب‌گاه‌های معتبر به سمت یک موضوع خاص، وب‌گاه‌ها را از تمام موضوعات انتخاب کردیم.

با توجه به این‌که برای ساخت مجموعه داده‌ای موردنظر، روش برچسب‌زنی در سطح میزبان را انتخاب نموده‌ایم، پس از جمع‌آوری آدرس صفحات فارسی مختلف، فهرست وب‌گاه‌های یکتا را از آن‌ها استخراج نمودیم.

خزش اولیه وب‌گاه‌ها

پس از فراهم کردن فهرستی از آدرس وب‌گاه‌های فارسی، می‌توان محتوای آن‌ها را با استفاده از یک خزشگر مناسب ذخیره نمود و در مراحل بعد از محتوای صفحات به صورت برون‌خط استفاده کرد. در این مرحله، پیدا کردن یک خزشگر مناسب برای وظیفه موردنظر از اهمیت ویژه‌ای برخوردار است. بدین منظور پس از بررسی قابلیت‌ها و ضعف‌های تعدادی از خزشگرهای موجود، در نهایت از دو خزشگر برای ساخت مجموعه داده‌ای موردنظر استفاده نمودیم. در ابتدا با توجه به قابلیت‌ها و واسط کاربری مناسبی که خزشگر Offline Explorer[1] دارد، از آن برای خزش فهرست اولیه وب‌گاه‌ها استفاده کردیم. این خزشگر این امکان را فراهم می‌کند که با یک بار اجرا و تنظیم سطح خزش وب‌گاه‌ها، تمام صفحات یک وب‌گاه را به طور کامل خزش کرده و به صورت برون‌خط از آن استفاده نماییم. در واقع این خزشگر یک مرورگر را شبیه‌سازی می‌کند که در آن می‌توان با تلیک کردن بر روی پیوندهای مختلف درون یک صفحه، به صفحات دیگر آن وب‌گاه و همچنین سایر صفحات مربوط به پیوندهای درون صفحه رفت. یکی از مزیت‌هایی که این خزشگر دارد، واسط کاربری مناسب و امکان انجام تنظیماتی مانند انتخاب و یا حذف بخش‌های مختلف صفحه از جمله عکس‌ها، فیلم و فایل‌های با فرمت خاص، در هنگام خزش می‌باشد.

پس از خزش وب‌گاه‌ها، به صورت برون‌خط و با استفاده از شبیه‌ساز مرورگر در Offline Explorer، یک بررسی اجمالی بر روی وب‌گاه‌های خزش شده انجام داده و وب‌گاه‌هایی را که شرایط موردنظر را ندارند، از فهرست حذف کردیم. برای مثال، یکی از مشکلاتی که در بررسی وب‌گاه‌ها به آن برخوردیم، سرعت از بین رفتن وب‌گاه‌های هرز می‌باشد. موتورهای جست‌وجو علاوه بر استفاده از انواع روش‌های شناسایی هرز وب و مقابله با صفحات هرز به صورت برخط، فهرست سیاهی از صفحات و وب‌گاه‌های هرز را نیز تهیه می‌کنند. با داشتن این فهرست در هنگام بازیابی نتایج، صفحات هرز درون این فهرست، توسط موتورهای جست‌وجو نمایه‌سازی نمی‌شوند. هرزنویسان برای فرار از این مشکل، معمولا پس از مدت زمان مشخصی، آدرس وب‌گاه‌های خود را تغییر می‌دهند که این امر باعث می‌شود بسیاری از نام‌های میزبان مربوط به وب‌گاه‌های هرز بدون استفاده باقی بماند. در نتیجه این نوع از وب‌گاه‌ها را که تعداد قابل توجهی داشتند،

[1] http://www.metaproducts.com/OE.html

از فهرست وب‌گاه‌های موردنظر حذف نمودیم.

جمع‌آوری وب‌گاه‌های هرز

پس از بررسی اجمالی وب‌گاه‌های خزش شده و پالایش آن‌ها، تعداد زیادی از وب‌گاه‌های هرز از میان فهرست وب‌گاه‌ها حذف شدند و تنها تعداد محدودی از آن‌ها باقی ماندند. بنابراین برای تکمیل مجموعه داده‌ای، همچنان به وب‌گاه‌های هرز بیشتری نیاز داشتیم. از طرف دیگر، با توجه به این‌که درصد وب‌گاه‌های هرز به نسبت وب‌گاه‌های معتبر بسیار کمتر می‌باشد، مشکل اصلی در ایجاد پیکره‌ای از وب‌گاه‌های هرز و معتبر، پیدا کردن وب‌گاه‌های هرز می‌باشد. بدین منظور از روش‌های دیگری برای جمع‌آوری این نوع از وب‌گاه‌ها استفاده کردیم که در ادامه به شرح آن‌ها می‌پردازیم.

در این مرحله، با در نظر گرفتن این امر که در میان پیوندهای درون یک صفحه هرز، معمولا تعدادی ارجاع به سایر صفحات هرز وجود دارد، با دنبال کردن مسیر پیوندهای خروجی از صفحات هرز، به دنبال جمع‌آوری صفحات هرز جدید پرداختیم. بدین منظور ابتدا فهرست آدرس‌های مجموعه وب‌گاه‌های هرز باقی‌مانده از مرحله قبل را به عنوان بذر اولیه به خزشگر Offline Explorer داده و صفحات را تا سه سطح خزش کردیم. با توجه به این‌که تمرکز ما در این مرحله بیشتر بر روی پیدا کردن وب‌گاه‌های هرز می‌باشد، بعد از خزش سطح اول، خزشگر را متوقف کرده و صفحات خزش شده جدید را بررسی نمودیم. از میان این صفحات، مجددا فهرست مربوط به وب‌گاه‌های یکتا و جدید را استخراج کرده و پس از بررسی اجمالی آن‌ها و حذف وب‌گاه‌های مسدود شده، وب‌گاه‌های هرز باقی‌مانده را به عنوان بذر به خزشگر دادیم. سپس فرآیند مرحله اول را مجددا تکرار نمودیم.

برای جلوگیری از سوگیری وب‌گاه‌های هرز به سمت یک مجموعه بذر مشخص، با استفاده از مجموعه کلیدواژه‌های رایج در پرس‌وجوها، به صورت تصادفی پرس‌وجوهای جدیدی ایجاد کرده و به موتورهای جست‌وجوی مختلف مانند گوگل و بینگ دادیم. نمونه‌هایی از این کلمات، «دانلود»، «عکس»، «بازی»، «آهنگ» و «فیلم» می‌باشد که طبق آمار گزارش شده توسط گوگل [۱]، به ترتیب بالاترین میزان استفاده را در

[۱] http://www.google.com/trends/

پرس‌وجوهای دو سال اخیر داشته‌اند. با توجه به این‌که محدود کردن پرس‌وجوها به این کلمات، نتایج را به سمت وب‌گاه‌های مشخص و محدودی متمایل می‌کند، از حدود صد کلمه رایج در پرس‌وجوها استفاده نمودیم و آن‌ها را با توجه به میزان ارتباطشان با یکدیگر به ده گروه تقسیم کردیم. سپس در هر گروه تعداد دو یا سه کلمه از آن را به صورت تصادفی کنار هم گذاشته و بدین ترتیب تعداد زیادی پرس‌وجو ایجاد نمودیم. پس از آن به ازای هر پرس‌وجو، ده نتیجه اول را بررسی کردیم. با توجه به این‌که صفحات هرز معمولا دارای تعداد زیادی کلیدواژه هستند، در میان نتایج بازیابی شده توسط موتور جست‌وجوی بینگ، به ازای هر پرس‌وجو به طور متوسط حدود دو الی سه صفحه هرز وجود داشت که البته بسیاری از آن‌ها مربوط به یک وب‌گاه مشترک بودند.

برچسب‌زنی وب‌گاه‌ها

پس از مشخص کردن فهرست نهایی آدرس وب‌گاه‌های خزش شده، برچسب وب‌گاه‌ها را مشخص نمودیم. برای این کار، با داشتن محتوای وب‌گاه‌ها به صورت برون‌خط، برای ساده‌سازی و افزایش سرعت برچسب‌زنی وب‌گاه‌ها، برنامه‌ای نوشتیم که دارای واسط گرافیکی با گزینه‌هایی برای مشخص کردن زبان صفحه، برچسب آن، نوع هرز بودن آن و اعلام تغییر خودکار آدرس صفحه به صفحه‌ای دیگر می‌باشد. با استفاده از این برنامه، می‌توان محتوای مجموعه‌ای از وب‌گاه‌ها را بررسی کرده و برچسب آن‌ها را مشخص نماییم.

خزش نهایی و ذخیره مجموعه وب‌گاه‌ها

پس از برچسب‌زنی نهایی وب‌گاه‌ها، برای خزش نهایی آن‌ها از HtmlUnit[1] که به زبان جاوا می‌باشد استفاده نمودیم. دلیل استفاده از این خزشگر، امکان خزش کردن بخش‌های پویای صفحه مانند کدهای جاوا اسکریپت و محتوای درون i-frameها می‌باشد. همچنین با استفاده از این خزشگر می‌توان هم‌زمان فرآیند هنجارسازی متن صفحه و جداسازی بخش‌های مختلف آن را انجام داد.

پس از انجام مراحل بالا، در نهایت ۳۰۰ وب‌گاه هرز فارسی و ۱۰۵۰ وب‌گاه معتبر فارسی جمع‌آوری،

[1] http://htmlunit.sourceforge.net/

خزش و نمایه‌سازی شدند. اطلاعات مربوط به خصوصیات آماری این مجموعه داده‌ای در ادامه در بخش ۱.۴ ارائه شده است.

۲.۱.۳ معرفی و تحلیل ویژگی‌های محتوایی بر روی وب‌گاه‌های فارسی

در این بخش، ابتدا به توضیح مختصر ویژگی‌های محتوایی که توسط Ntoulas و همکاران [۸] معرفی شده است می‌پردازیم. سپس با معرفی ویژگی‌های محتوایی ارائه شده در [۲۱، ۲۹، ۱۰۰] رفتار وب‌گاه‌های هرز فارسی را بر اساس مقادیر مختلف هر یک از این ویژگی‌ها تحلیل می‌کنیم. در نهایت برای بهبود کارایی رده‌بند نهایی، تعدادی ویژگی جدید معرفی می‌نماییم.

ویژگی‌های گروه ۱: ویژگی‌های پایه

پس از ساخت مجموعه داده‌ای برچسب خورده شامل وب‌گاه‌های فارسی، در مرحله اول، ویژگی‌های محتوایی معرفی شده در مقاله [۸] را به عنوان ویژگی‌های پایه[۱]، از وب‌گاه‌های فارسی استخراج کرده و میزان تاثیر آن‌ها را در شناسایی وب‌گاه‌های هرز فارسی بررسی نمودیم. این ویژگی‌ها عبارتند از:

- تعداد کلمات هر صفحه: یکی از روش‌های رایج در میان هرزنویسان، روش انباشتگی کلیدواژه‌ها است. در این روش با استفاده از انواع کلیدواژه‌ها و تکرار آن‌ها درون صفحات خود سعی می‌کنند رتبه صفحه خود را به ازای پرس‌وجوهای بیشتری بالا ببرند. بنابراین تعداد کلمات در این نوع از صفحات هرز به طور متوسط از صفحات معتبر بیشتر می‌باشد.
- تعداد کلمات عنوان هر صفحه: با توجه به این‌که برخی از موتورهای جست‌وجو به کلمات درون عنوان صفحات وزن بیشتری اختصاص می‌دهند، در بسیاری از صفحات هرز از روش انباشتگی کلیدواژه‌ها در عنوان صفحه استفاده می‌شود.
- متوسط طول کلمات موجود در صفحه: یکی از روش‌های دیگری که هرزنویسان برای بالا بردن رتبه

[۱] baseline

صفحه خود به‌کار می‌برند، استفاده از کلمات ترکیبی است که از اتصال دو تا سه کلمه به یکدیگر ایجاد می‌شوند. با استفاده از این روش هرزنویسان می‌توانند رتبه صفحه خود را به ازای پرس‌وجوهایی که کاربران فراموش می‌کنند بین کلمات آن از فاصله استفاده کنند بالا ببرند. بنابراین انتظار می‌رود که در این نوع از صفحات هرز، متوسط طول کلمات بیشتر از صفحات معتبر باشد.

- درصد متن پیوند درون هر صفحه: هرزنویسان برای بالا بردن رتبه صفحات خود تعداد زیادی صفحه ایجاد می‌کنند که به صفحات موردنظر ارجاع می‌دهند. این صفحات معمولا محتوای زیادی ندارند و دارای تعداد زیادی پیوند هستند که هر کدام به صفحاتی با موضوعات مختلف اشاره می‌کنند. بنابراین در این‌گونه صفحات هرز، درصد بیشتری از صفحه با استفاده از متن پیوند پر شده است که معمولا دارای کلیدواژه‌ها با موضوعات مختلف می‌باشند.

- درصد محتوای قابل مشاهده در هر صفحه: برخی از موتورهای جست‌وجو برای پیدا کردن نتایج مرتبط با پرس‌وجوهای کاربران، علاوه بر متن اصلی صفحات، بخش‌هایی از محتوای صفحات را نیز که درون ابربرچسب‌های HTML می‌باشد، نمایه‌سازی می‌کنند. ابربرچسب‌های درون بخش header صفحه و برچسب ALT مربوط به تصاویر درون صفحه، که دارای توضیحاتی پیرامون تصاویر مربوطه است، نمونه‌هایی از این بخش‌ها می‌باشند. با توجه به این رفتار موتورهای جست‌وجو، هرزنویسان برای بالا بردن رتبه صفحات هرز خود، از روش انباشتگی کلیدواژه‌ها در این بخش‌ها استفاده می‌کنند.

- درصد فشرده‌سازی: با توجه به این‌که برخی از موتورهای جست‌وجو به صفحاتی که تعداد تکرار کلمات پرس‌وجو در آن‌ها بیشتر باشد وزن بیشتری می‌دهند، هرزنویسان معمولا کلیدواژه‌ها را به تعداد زیاد درون صفحات خود استفاده می‌کنند. با استفاده از الگوریتم‌های فشرده‌سازی مانند GZIP [۱۰۱] می‌توان درصد تکراری بودن محتوای صفحات را محاسبه کرد. برای محاسبه این ویژگی ابتدا هر صفحه را با استفاده از الگوریتم GZIP فشرده کردیم. سپس اندازه صفحه را بر اندازه صفحه پس از فشرده‌سازی تقسیم نمودیم. هر چقدر مقدار این تقسیم بزرگتر باشد، بدین معنا می‌باشد که صفحه موردنظر، قابلیت فشرده‌سازی بیشتری دارد؛ به عبارت دیگر، محتوای تکراری در آن صفحه بیشتر بوده و در نتیجه احتمال هرز بودن آن صفحه نیز بیشتر می‌باشد.

- درصدی از صفحه که شامل کلمات مشهور می‌باشد: برای محاسبه این ویژگی، کلمات پرتکرار موجود در مجموعه وب‌گاه‌ها استخراج شده و n کلمه اول به عنوان مجموعه کلمات مشهور در نظر گرفته می‌شود. سپس به ازای هر صفحه، تعداد کلمات مشهور درون صفحه را، بر تعداد کل کلمات موجود در آن تقسیم می‌نماییم. انتظار می‌رود که با استفاده از این ویژگی بتوان وب‌گاه‌های هرزی را که به صورت تصادفی تعدادی از کلیدواژه‌ها را درون صفحات خود تکرار کرده‌اند، شناسایی کرد. در این‌گونه از وب‌گاه‌ها، معمولا درصد کلمات مشهور مانند ایست‌واژه‌ها، کمتر از درصد این کلمات در وب‌گاه‌های معتبر است.

- درصدی از کلمات مشهور که در صفحه استفاده شده است: این ویژگی مکمل ویژگی قبلی است و نشان می‌دهد که چند درصد از کلمات مشهوری که در متن استفاده شده‌اند از یکدیگر متمایز هستند. برای محاسبه این ویژگی، تعداد کلماتی از مجموعه کلمات مشهور، که در متن آمده است را بر n، که همان تعداد کل کلمات مشهور می‌باشد، تقسیم می‌نماییم. با استفاده از این ویژگی می‌توان وب‌گاه‌های هرزی را شناسایی کرد که برای فریب ویژگی قبل، یک یا چند کلمه مشهور را به تعداد زیاد درون صفحات خود تکرار می‌کنند. تعداد کلمات مشهور یکتا در این‌گونه از وب‌گاه‌های هرز، نسبت به وب‌گاه‌های معتبر کمتر می‌باشد.

- احتمال شباهت nـ گرام‌های مستقل[۱]: این ویژگی میزان تکراری بودن محتوای صفحات را در سطح nـ گرام محاسبه می‌کند. با استفاده از این ویژگی می‌توان دو نوع از وب‌گاه‌های هرز را شناسایی کرد. نوع اول وب‌گاه‌هایی هستند که یک بخش از محتوای آن‌ها چندین بار درون صفحه تکرار شده است. دسته دوم وب‌گاه‌هایی هستند که با کپی کردن بخش‌های مختلف از صفحات مختلف دارای محتوای نامرتبط با یکدیگر می‌باشند. بنابراین، در صورتی که برای مقادیر این ویژگی دو حد آستانه حداقل و حداکثر تعریف نماییم، احتمال هرز بودن صفحات برای صفحات خارج از این بازه، بیشتر

[۱] independent n-gram likelihood

از صفحات داخل این بازه خواهد بود. این ویژگی به صورت زیر محاسبه می‌شود:

$$independent\_likelihoods = -\frac{1}{k}\sum_{i=0}^{(k-1)} \log P(w_{i+1}, ..., w_{i+n}) \qquad (۱.۳)$$

که در آن $P(w_{i+1}, ..., w_{i+n})$ احتمال $(i+1)$امین nـ گرام است که با تقسیم تعداد رخداد این nـ گرام به تعداد کل nـ گرام‌ها محاسبه می‌شود.

- احتمال شباهت nـ گرام‌های شرطی[۱]: این ویژگی مانند ویژگی قبل است، با این تفاوت که برای محاسبه ویژگی قبل، nـ گرام‌ها به طور مستقل از هم در نظر گرفته می‌شوند. در صورتی که برای محاسبه این ویژگی، احتمال هر nـ گرام به شرط وجود $(n-1)$ کلمه اول آن محاسبه می‌شود. نحوه محاسبه این ویژگی به صورت زیر است:

$$conditional\_likelihoods = -\frac{1}{k}\sum_{i=0}^{(k-1)} \log P(w_n | w_{i+1}, ..., w_{i+n-1}) \qquad (۲.۳)$$

به طوری‌که؛

$$P(w_n | w_{i+1}, ..., w_{i+n-1}) = \frac{P(w_{i+1}, ..., w_{i+n})}{P(w_{i+1}, ..., w_{i+n-1})} \qquad (۳.۳)$$

ویژگی‌های گروه ۲: ویژگی‌های مکمل

در مرحله بعد، برای شناسایی انواع بیشتری از وب‌گاه‌های هرز فارسی، تعدادی ویژگی محتوایی دیگر [۲۱، ۲۹، ۱۰۰] را از این صفحات استخراج کرده و میزان کارایی و تاثیر آن‌ها را در شناسایی وب‌گاه‌های هرز بررسی نمودیم. در ادامه هر یک از این ویژگی‌ها را به طور مختصر توضیح می‌دهیم.

- تعداد عکس‌ها: با بررسی تعداد عکس‌های درون وب‌گاه‌های فارسی به نظر می‌رسد که هر چقدر تعداد عکس‌ها، درون یک صفحه بیشتر باشد، احتمال هرز بودن آن صفحه بیشتر است. این امر

[۱] conditional n-gram likelihood

می‌تواند به این دلیل باشد که وب‌گاه‌های هرز معمولا با اهداف تبلیغاتی و تجاری ایجاد می‌شوند. استفاده از تصاویر تبلیغاتی مختلف در این‌گونه از صفحات می‌تواند در جذب مشتری تاثیر بسزایی داشته باشد. رفتار وب‌گاه‌های فارسی در برابر این ویژگی، برخلاف رفتاری است که توسط Prieto و همکاران [۲۹] برای وب‌گاه‌های انگلیسی بیان شده است. نظر آن‌ها این است که با توجه به این‌که وب‌گاه‌های هرز اکثرا به صورت خودکار ایجاد شده و با مطالب تصادفی پر می‌شوند، تعداد تصاویر موجود در آن‌ها کم می‌باشد.

- طول URL: هر URL دارای یک نام میزبان و نام دامنه است. برخلاف نام دامنه که محدود به تعداد مشخصی دامنه شناخته شده است، هر اسمی می‌تواند به عنوان نام میزبان انتخاب شود. برای انتخاب نام میزبان معمولا از اسامی مرتبط با موضوع صفحه، از عنوان صفحه و یا از مدل اختصاری آن استفاده می‌کنند. با توجه به این‌که برخی از موتورهای جست‌وجو، URL وب‌گاه‌ها را نیز نمایه‌سازی می‌کنند، بسیاری از هرزنویسان، برای افزایش امتیاز صفحات هرز خود، از کلیدواژه‌های زیادی درون URL صفحات خود استفاده می‌کنند. این کلمات معمولا با استفاده از علامت‌های نگارشی مختلف، مانند خط فاصله از یکدیگر جدا شده‌اند. با در نظر گرفتن این مهم، انتظار می‌رود به طور متوسط طول URL وب‌گاه‌های هرز فارسی نسبت به وب‌گاه‌های معتبر بیشتر باشد.

- تعداد کلمات درون ابربرچسب کلیدواژه‌ها و توضیحات صفحه: در هر صفحه HTML سه بخش عنوان، ابربرچسب کلیدواژه‌ها و ابربرچسب توضیحات وجود دارد که یک موتور جست‌وجو برای اطلاع از این‌که یک صفحه پیرامون چه موضوع یا موضوعاتی است، از محتوای این بخش‌ها استفاده می‌کند. در نتیجه این بخش از صفحات مکان خوبی برای هرزنویسان است که با استفاده از روش انباشتگی کلیدواژه‌ها در آن‌ها، رتبه صفحه هرز خود را افزایش دهند. بررسی وب‌گاه‌های فارسی نشان می‌دهد که متوسط تعداد کلمات موجود در ابربرچسب کلیدواژه‌ها و توضیحات، برای وب‌گاه‌های هرز بیشتر از وب‌گاه‌های معتبر است.

- تعداد کلمات درون متن پیوند: در میان صفحات وب، صفحاتی هستند که دارای پیوندهای زیاد به صفحات مختلف به همراه توضیحاتی پیرامون آن صفحات هستند. با توجه به این‌که موتورهای

جست‌وجو برای محاسبه امتیاز صفحات وب از کلمات درون متن پیوند مربوط به پیوندهای آن‌ها، که در سایر صفحات وجود دارد نیز استفاده می‌کنند، هرزنویسان برای افزایش رتبه صفحات درون وب‌گاه خود، صفحاتی را ایجاد می‌نمایند که دارای تعداد زیادی ارجاع به صفحات موردنظر، به همراه متن پیوند مربوط به آن‌ها می‌باشند. محاسبه این ویژگی برای وب‌گاه‌های فارسی نشان می‌دهد که این امر در میان وب‌گاه‌های فارسی، امری رایج می‌باشد.

- شباهت کسینوسی[۱] [۱۰۲] بین بخش‌های مختلف صفحه: هر صفحه وب را می‌توان به سه بخش اصلی عنوان، بدنه و متن پیوند تقسیم کرد. شباهت کسینوسی معیاری است که میزان شباهت دو متن را بر اساس روش‌های وزن‌دهی مختلف محاسبه می‌نماید. در نتیجه با استفاده از این معیار می‌توان میزان شباهت بین محتوای این سه بخش اصلی را محاسبه کرد. در این پژوهش، برای محاسبه این ویژگی از دو روش وزن‌دهی دودویی[۲] و TF-IDF استفاده کرده و شباهت کسینوسی بین هر زوج بخش اصلی را محاسبه نمودیم. بررسی رفتار وب‌گاه‌های فارسی نسبت به مقادیر مختلف این ویژگی نشان می‌دهد که احتمال هرز بودن وب‌گاه‌ها برای مقادیر کمتر از یک آستانه حداقلی و بیشتر از یک آستانه حداکثری افزایش می‌یابد. دلیل افزایش احتمال هرز بودن صفحات به ازای کاهش بیش از حد مقدار این ویژگی‌ها را می‌توان وجود صفحات هرزی دانست که با رونوشت بخش‌های مختلف از محتویات صفحات مختلف سعی دارند امتیاز صفحه خود را افزایش دهند. همچنین صفحاتی که مقدار این ویژگی برای آن‌ها از یک آستانه مشخص بیشتر است، صفحات هرزی هستند که برای افزایش رتبه صفحه خود از روش انباشتگی کلیدواژه‌ها استفاده کرده‌اند.

- تعداد ایست‌واژه‌ها درون هر صفحه: یکی از روش‌های نمایه‌سازی رایج در موتورهای جست‌وجو، حذف ایست‌واژه‌ها قبل از نمایه‌سازی صفحات می‌باشد. بنابراین هرزنویسان با حذف ایست‌واژه‌ها از وب‌گاه‌های خود و با استفاده از قرار دادن کلیدواژه‌های تصادفی در آن‌ها، امتیاز وب‌گاه خود را به ازای پرس‌وجوهای زیادی بالا می‌برند. بنابراین انتظار می‌رود با استفاده از این ویژگی بتوان وب‌گاه‌های هرزی را که با این روش تولید می‌شوند شناسایی کرد. اما همان‌طور که در مقاله [۲۹]

[۱] cosine similarity
[۲] binary

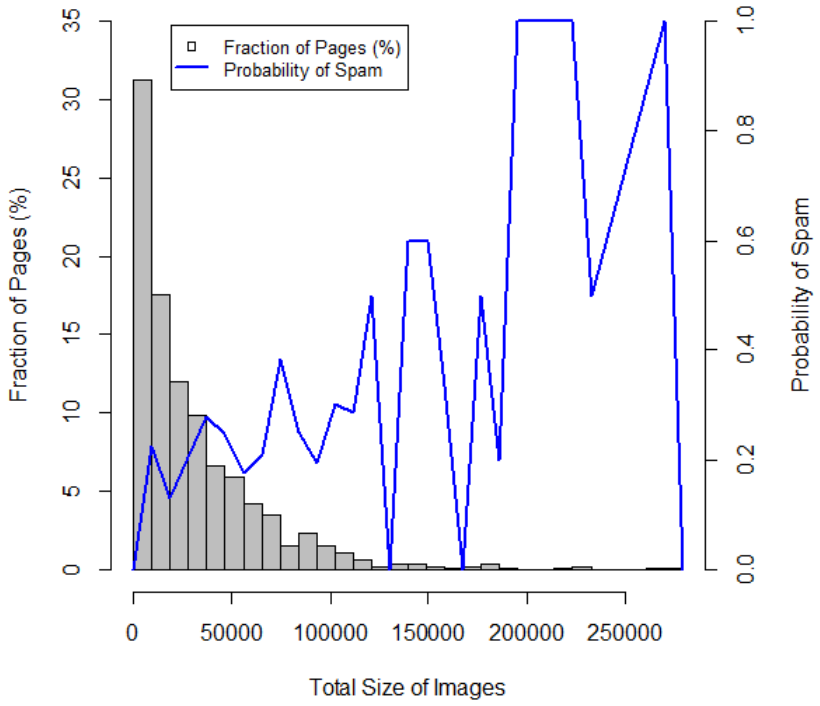
نشان داده می‌شود که این ویژگی تاثیر چندانی در شناسایی این دسته از وب‌گاه‌های هرز ندارد، با بررسی وب‌گاه‌های هرز فارسی نیز به چنین نتیجه‌ای رسیدیم. برای رفع این مشکل، در ادامه در بخش ویژگی‌های جدید، مدل بهبود یافته‌ای از این ویژگی را با عنوان «درصدی از صفحه که از ایست‌واژه‌ها تشکیل شده‌است»، معرفی می‌نماییم.

- تعداد پیوندهای خروجی: برخی از صفحات هرز، فقط با هدف افزایش PageRank سایر صفحات هرز ایجاد می‌شوند. این صفحات معمولا بخشی از یک دهکده پیوندی هستند که با دادن ارجاعات زیاد به سایر صفحات هرز، که معمولا مربوط به یک وب‌گاه مشترک می‌باشند، سعی بر افزایش رتبه آن صفحات و درنتیجه افزایش رتبه صفحه خود را دارند. پس از تحلیل این ویژگی بر روی وب‌گاه‌های فارسی به این نتیجه رسیدیم که احتمال هرز بودن یک وب‌گاه با افزایش تعداد پیوندهای خروجی از آن وب‌گاه افزایش می‌یابد.

ویژگی‌های گروه ۳: ویژگی‌های جدید

در این مرحله برای بهبود رده‌بندی وب‌گاه‌های فارسی تعدادی ویژگی جدید معرفی می‌نماییم. ابتدا این ویژگی‌ها را از مجموعه داده‌ای فارسی استخراج کرده و پس از رسم نمودار توزیع وب‌گاه‌ها برای هر یک از ویژگی‌ها به طور مجزا و تحلیل هر یک از آن‌ها، میزان تاثیر آن‌ها را در نتیجه نهایی رده‌بندی محاسبه کردیم. این ویژگی‌ها به شرح زیر می‌باشند:

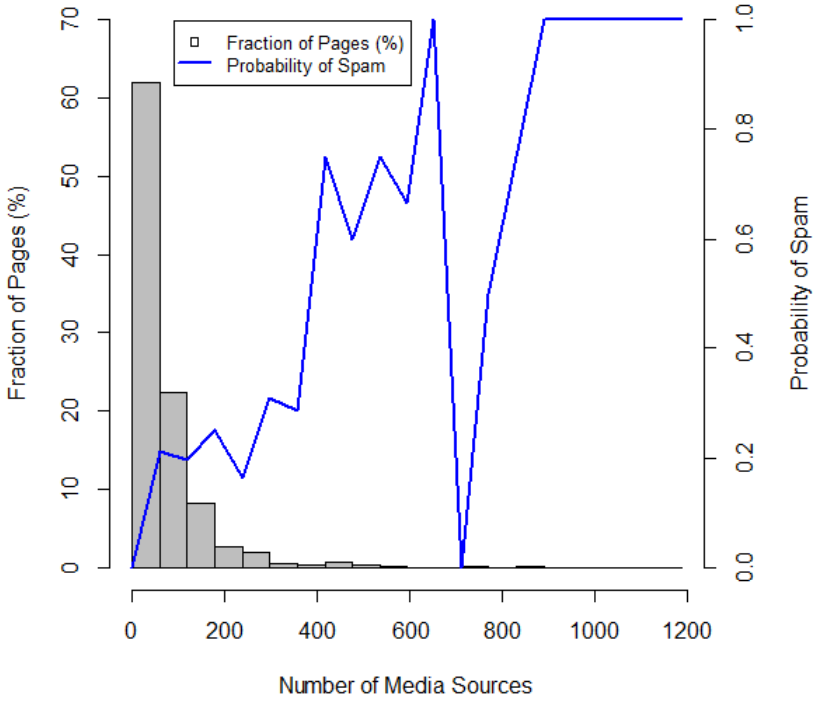
- مجموع اندازه عکس‌های درون هر صفحه: بررسی تعداد زیادی از وب‌گاه‌های فارسی نشان می‌دهد که به طور متوسط، تعداد تصاویر و مجموع اندازه آن‌ها در وب‌گاه‌های هرز، بیشتر از وب‌گاه‌های معتبر می‌باشد. در این‌گونه از وب‌گاه‌های هرز، درصد محتوای متنی درون صفحه، کمتر از سایر وب‌گاه‌ها می‌باشد. هدف از این امر، تبلیغ محصولات و جذب مشتری و رسیدن به سایر اهداف تجاری می‌باشد. بزرگ بودن تصاویر، علاوه بر این‌که در جذب مشتری تاثیر قابل توجهی دارد، احتمال تلیک کردن بر روی آن تصاویر و رفتن به صفحات تبلیغاتی مربوط به آن را افزایش می‌دهد. بنابراین علاوه بر تعداد عکس‌های درون هر صفحه، مجموع اندازه عکس‌هایی که در هر صفحه نمایش داده می‌شود



شکل ۴.۳: رفتار وب‌گاه‌های فارسی به ازای مقادیر مختلف ویژگی «مجموع اندازه عکس‌های درون هر صفحه»

را به عنوان یک ویژگی جدید به سایر ویژگی‌ها اضافه کردیم. نمودار ۴.۳ توزیع وب‌گاه‌های فارسی و همچنین احتمال هرز بودن این وب‌گاه‌ها را به ازای مقادیر مختلف این ویژگی نشان می‌دهد. با توجه به این نمودار، با افزایش مجموع اندازه عکس‌های درون هر وب‌گاه، احتمال هرز بودن آن وب‌گاه نیز افزایش می‌یابد.

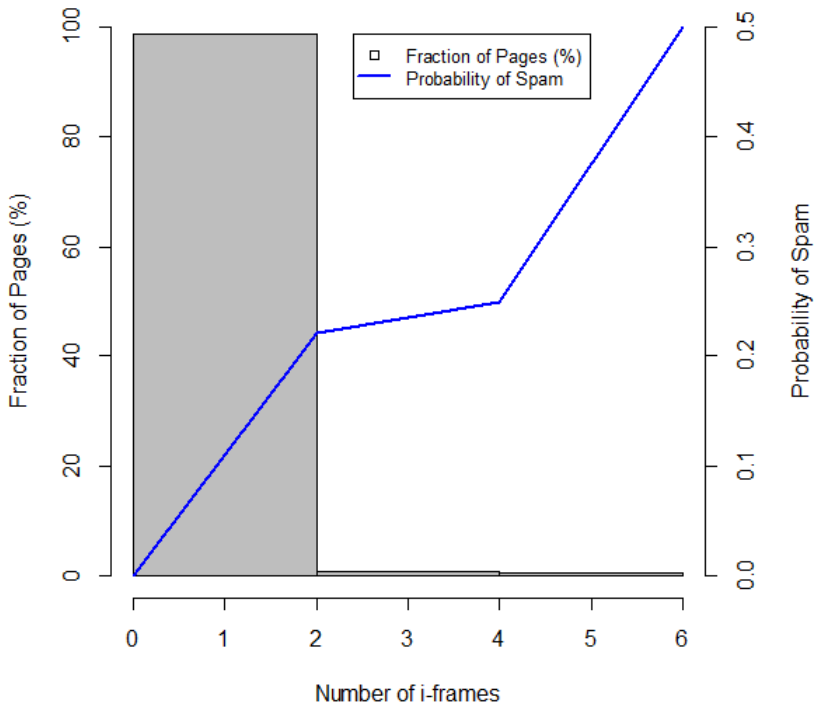
- تعداد منابع چندرسانه‌ای درون هر صفحه: خصوصیت دیگری که در زمان بررسی وب‌گاه‌های هرز فارسی مشاهده می‌نماییم، استفاده هرزنویسان از منابع چند رسانه‌ای برای جذب کاربران به سمت صفحات خود و یا سایر وب‌گاه‌های هرز می‌باشد. برای مثال یک کاربر با ورود به یک صفحه هرز که محتوای مفیدی ندارد اما به صورت خودکار موسیقی زیبایی را پخش می‌نماید، برای گوش دادن به موسیقی مدت زمان بیشتر را در آن صفحه می‌ماند. با توجه به نمودار ۵.۳ مشاهده می‌نماییم که با افزایش مقدار این ویژگی، احتمال هرز بودن وب‌گاه‌های فارسی در کل یک روند صعودی دارد. همان‌طور که در نمودار مشخص است، به ازای مقدار ۷۱۲ برای تعداد منابع چندرسانه‌ای، میزان احتمال هرز بودن وب‌گاه‌ها صفر می‌باشد. دلیل این امر کوچک بودن مجموعه داده‌ای فارسی در



شکل ۵.۳: رفتار وب‌گاه‌های فارسی به ازای مقادیر مختلف ویژگی «تعداد منابع چندرسانه‌ای»

مقایسه با تعداد کل صفحات وب می‌باشد که باعث می‌شود پراکندگی وب‌گاه‌ها بر روی مقادیر مختلف یک ویژگی یکنواخت نباشد. در نتیجه با توجه به این‌که در این بازه هیچ وب‌گاهی با این تعداد منابع چندرسانه‌ای وجود ندارد، بدیهی است که تعداد وب‌گاه‌های هرز و به دنبال آن، احتمال هرز بودن وب‌گاه‌ها نیز صفر می‌باشد.

- تعداد i-frameهای درون هر صفحه: استفاده از i-frameها در وب‌گاه‌های هرز فارسی امری رایج است که به هرزنویسان کمک می‌کند تا بتوانند بخش‌هایی از صفحه یا صفحات دیگر را درون صفحه خود نمایش دهند. این صفحات معمولا با استفاده از تعداد زیادی کلیدواژه درون ابربرچسب یا به صورت پنهان درون بدنه صفحه، کاربران را به سمت صفحه خود هدایت می‌کنند. سپس با استفاده از مطالب نامرتبط موجود در صفحات دیگر، مدت زمان ماندن کاربران را در صفحه خود افزایش می‌دهند. همچنین تعدادی از این صفحات، رونوشتی از پیوندهای موجود در صفحات دیگر را درون i-frameها قرار داده و با استفاده از این پیوندها کاربران را به سمت سایر صفحات هرز هدایت می‌کنند. نمودار ۶.۳ نشان می‌دهد که صفحاتی که از i-frameهای بیشتری استفاده کرده‌اند با احتمال

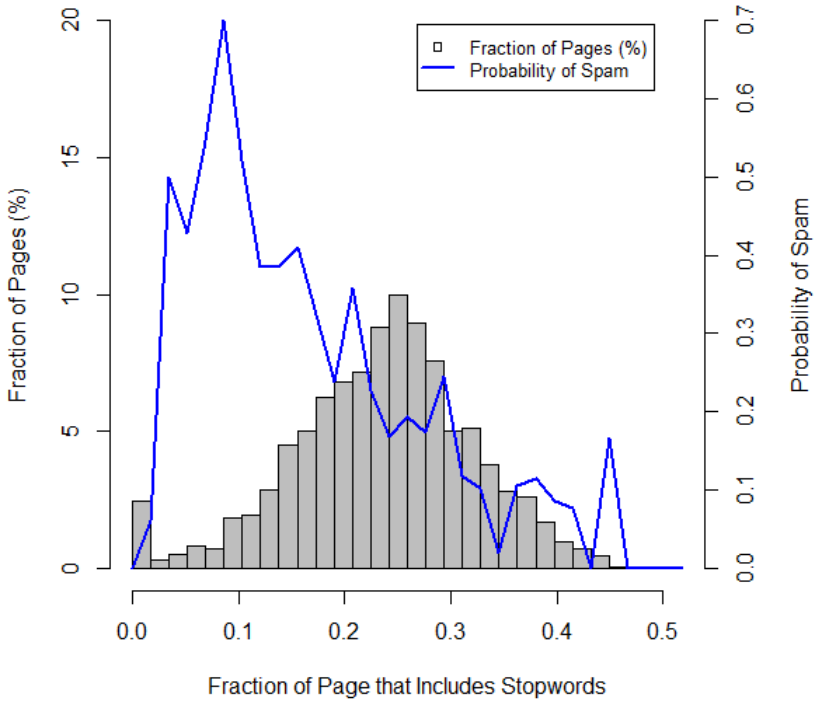


شکل ۶.۳: رفتار وب‌گاه‌های فارسی به ازای مقادیر مختلف ویژگی «تعداد i-frameها»

بیشتری هرز می‌باشند.

- درصدی از صفحه که از ایست‌واژه‌ها تشکیل شده است: این ویژگی در واقع مدل بهبود یافته‌ای از ویژگی «تعداد ایست‌واژه‌ها در هر صفحه» می‌باشد. همان‌طور که در بخش مربوط به مجموعه ویژگی‌های مکمل توضیح داده شد، به دلیل اختلاف طول صفحات با یکدیگر، محاسبه تعداد کل ایست‌واژه‌های هر صفحه برای شناسایی وب‌گاه‌های هرز کاربرد چندانی ندارد. بنابراین در این پژوهش ویژگی دیگری را ارائه می‌دهیم که برای محاسبه آن تعداد ایست‌واژه‌ها را بر تعداد کل کلمات موجود در صفحه تقسیم می‌نماییم. این امر باعث می‌شود خطاهایی که به دلیل یکسان نبودن طول صفحات وب ایجاد می‌شود تا حدود زیادی کاهش یابد. با توجه به نمودار ۷.۳ مشاهده می‌کنیم که احتمال هرز بودن وب‌گاه‌های فارسی، با کاهش تعداد ایست‌واژه‌ها افزایش می‌یابد. دلیل این امر وجود وب‌گاه‌های هرزی است که با استفاده از تعداد زیادی کلیدواژه بدون وجود جمله‌های معنادار ایجاد شده‌اند.

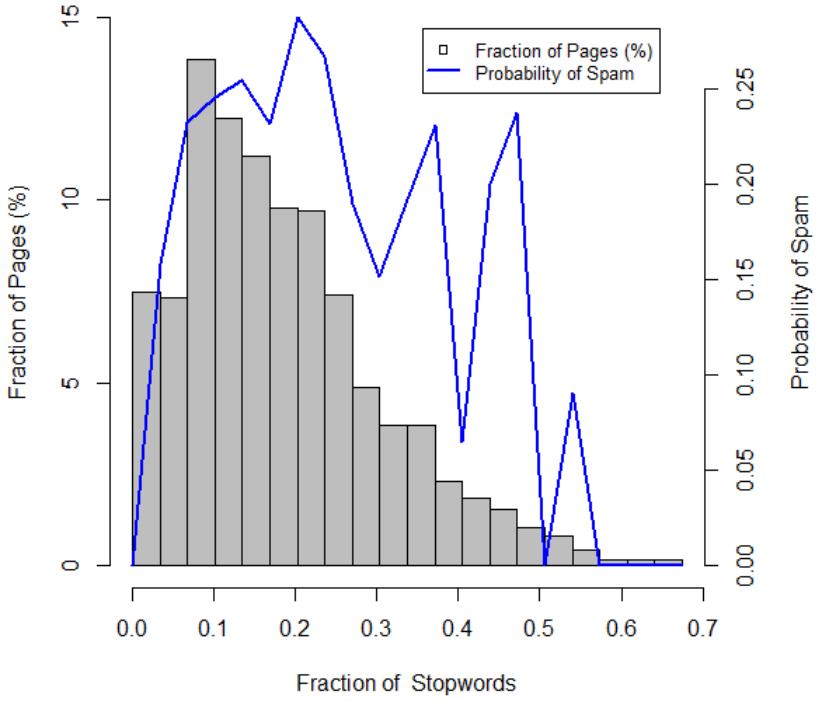
- درصدی از ایست‌واژه‌ها که درون صفحه استفاده شده است: این ویژگی به عنوان مکملی برای ویژگی



شکل ۷.۳: رفتار وب‌گاه‌های فارسی به ازای مقادیر مختلف ویژگی «درصدی از صفحه که شامل ایست‌واژه است»

قبل می‌باشد. هرزنویسان می‌توانند برای فریب روش قبل، تعداد کمی از ایست‌واژه‌ها را به دفعات زیاد درون صفحه خود تکرار کنند. همچنین وجود این ویژگی باعث می‌شود آن بخش از خطاهایی که به علت وجود تفاوت در طول صفحات ایجاد می‌شود از بین برود. برای مثال در صورتی که یک صفحه وب تنها از سه کلمه تشکیل شده باشد و دو کلمه آن جزء ایست‌واژه‌ها باشد، مقدار ویژگی قبل برای این صفحه ۶۷٪ محاسبه می‌شود. در صورتی که درصد ایست‌واژه‌های یکتا در این صفحه نشان می‌دهد که تعداد ایست‌واژه‌های یکتای موجود در این صفحه زیاد نمی‌باشد. نمودار ۸.۳ نمایانگر نحوه توزیع وب‌گاه‌های فارسی بر روی مقادیر مختلف این ویژگی و همچنین احتمال هرز بودن این وب‌گاه‌ها می‌باشد. با توجه به نمودار، وب‌گاه‌هایی که تعداد ایست‌واژه‌های یکتا در آن‌ها کمتر است، با احتمال بیشتری هرز می‌باشند.

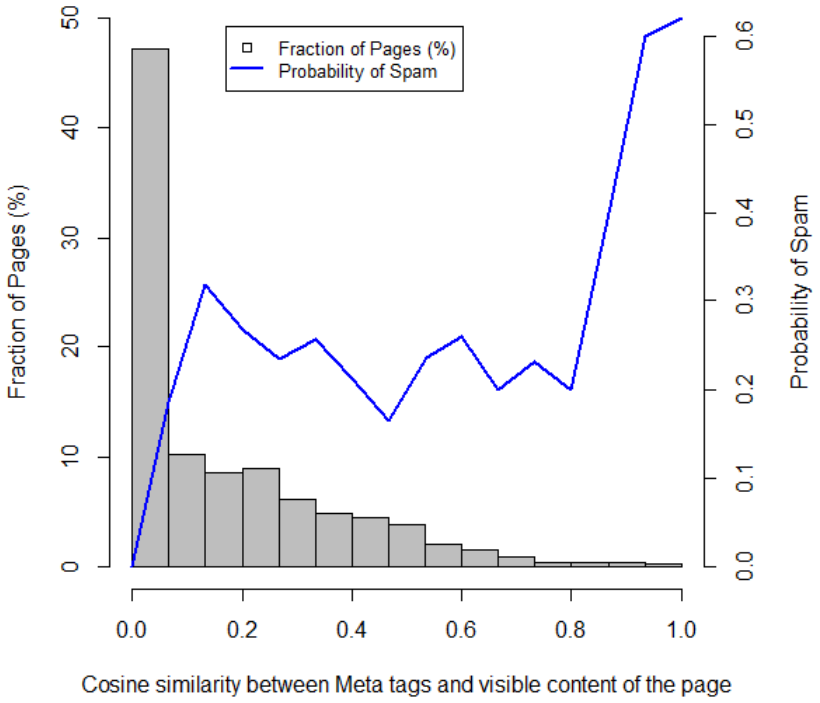
- شباهت کسینوسی بین ابربرچسب‌ها و محتوای قابل مشاهده هر صفحه: با توجه به این‌که برخی از موتورهای جست‌وجو به کلمات درون ابربرچسب‌ها وزن زیادی می‌دهند، بسیاری از هرزنویسان این بخش از صفحات خود را با تعداد زیادی کلیدواژه پر می‌کنند، که در این میان بسیاری از آن‌ها با موضوع



شکل ۸.۳: رفتار وب‌گاه‌های فارسی به ازای مقادیر مختلف ویژگی «درصد ایست‌واژه‌ها»

و محتوای صفحه نامرتبط هستند. با محاسبه شباهت کسینوسی بین این ابربرچسب‌ها و محتوای قابل مشاهده توسط کاربر و در نظر گرفتن این معیار به عنوان یک ویژگی در رده‌بندی وب‌گاه‌ها می‌توانیم این نوع از وب‌گاه‌های هرز را تشخیص دهیم. همان‌طور که در نمودار ۹.۳ مشاهده می‌نماییم، مقدار شباهت کسینوسی در صفحات معتبر حدود ۰/۵ بوده و به هر میزان که مقدار این معیار از ۰/۵ کمتر یا بیشتر می‌شود، احتمال هرز بودن وب‌گاه‌های فارسی نیز افزایش می‌یابد. کاهش این شباهت نشان‌گر وب‌گاه‌های هرزی است که برای مرتبط کردن محتوای وب‌گاه خود با تعداد زیادی از پرس‌وجوها، از کلیدواژه‌های مختلفی در صفحاتشان استفاده می‌کنند. همچنین افزایش این شباهت نشان‌دهنده وب‌گاه‌های هرزی است که برای بالا بردن وزن خود به ازای پرس‌وجوهای مشخص، مجموعه‌ای از کلیدواژه‌های مرتبط را به تعداد زیاد درون صفحات خود تکرار می‌کنند.

- میزان ابربرچسب‌های جاوا اسکریپت در هر صفحه: کدهای جاوا اسکریپت معمولا زمانی اجرا می‌شوند که رخدادی به وقوع می‌پیوندد. هرزنویسان معمولا با استفاده از کدهای جاوا اسکریپت، برخی رخدادهای پویا را کنترل می‌کنند. برای مثال زمانی که وارد یک صفحه هرز می‌شویم یا زمانی

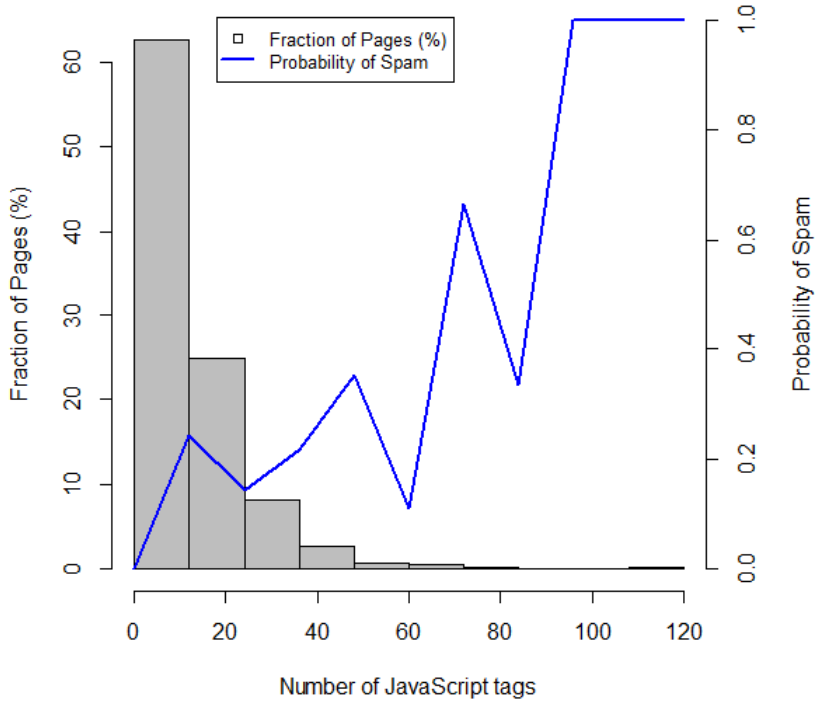


شکل ۹.۳: رفتار وب‌گاه‌های فارسی به ازای مقادیر مختلف «شباهت کسینوسی بین ابربرچسب‌ها و محتوای قابل مشاهده صفحه»

که روی یک پیوند درون صفحه تلیک می‌نماییم، تعدادی صفحه (معمولا تبلیغاتی) به طور خودکار توسط مرورگر نمایش داده می‌شود. در برخی مواقع نیز قبل از نمایش صفحه، کاربران به طور خودکار به صفحه دیگری هدایت می‌شوند. نمودار ۱۰.۳ رشد صعودی احتمال هرز بودن صفحات را به ازای افزایش مقدار این ویژگی نشان می‌دهد.

انتخاب ویژگی و رده‌بندی

پس از استخراج هر دسته از ویژگی‌ها و بررسی رفتار آن‌ها در شناسایی هرز وب فارسی، برای مشخص کردن بهترین ویژگی‌ها و بهبود دقت رده‌بندی، از روش انتخاب ویژگی χ^2 -test استفاده نمودیم. در این روش میزان وابستگی هر یک از ویژگی‌ها با دو کلاس هرز و معتبر محاسبه شده و سپس ویژگی‌ها بر اساس امتیازی که بدست می‌آورند به صورت نزولی مرتب می‌شوند. هر چقدر یک ویژگی امتیاز بیشتری بدست آورد، بدین معنا می‌باشد که میزان وابستگی آن ویژگی به کلاس‌های تعریف شده بیشتر است و درنتیجه انتظار می‌رود که آن ویژگی بتواند کلاس نمونه‌ها را درست‌تر تشخیص دهد. در نهایت با استفاده از روش حذف



شکل ۱۰.۳: رفتار وب‌گاه‌های فارسی به ازای مقادیر مختلف ویژگی «میزان ابربرچسب‌های جاوا اسکریپت»

پس‌رو، مجموعه‌ای از ویژگی‌ها را به عنوان ویژگی‌های بهینه انتخاب نمودیم. در این روش ابتدا با استفاده از تمام ویژگی‌های استخراج شده وب‌گاه‌ها را رده‌بندی کرده، سپس در هر مرحله ویژگی با کمترین امتیاز را حذف می‌نماییم. در صورتی که حذف یک ویژگی باعث کاهش کارایی رده‌بندی شود، آن ویژگی را مجددا به مجموعه ویژگی‌های بهینه اضافه می‌نماییم. پس از حذف ویژگی‌های غیربهینه، برای پیش‌بینی برچسب وب‌گاه‌ها، انواع الگوریتم‌های یادگیری ماشین مانند درخت تصمیم C4.5، جنگل تصادفی، شبکه‌های عصبی، الگوریتم‌های بیزین، SVM و k نزدیک‌ترین همسایه را بررسی کرده و در نهایت، روش جنگل تصادفی را به دلیل کارایی بالاتر آن برای شناسایی وب‌گاه‌های هرز انتخاب نمودیم. همچنین، برای بررسی میزان تاثیر هر ویژگی به طور مجزا در کارایی رده‌بند، فرآیند رده‌بندی را با استفاده از هر یک از ویژگی‌های بهینه به صورت مجزا انجام داده و دقت شناسایی وب‌گاه‌های هرز را برای آن محاسبه نمودیم.

۳.۱.۳ ارائه یک سامانه شناساگر هرز وب فارسی به نام PSD-SYS

در این بخش یک سامانه جدید به نام PSD-SYS[۱] را ارائه می‌دهیم که در آن از مدل جدیدی به نام BOSW برای استخراج ویژگی‌ها و همچنین روش انتخاب ویژگی بر اساس تعداد رخداد کلمات استفاده می‌شود. در این سامانه از الگوریتم SVM برای رده‌بندی وب‌گاه‌ها استفاده می‌شود. در ادامه به توضیح دقیق‌تر درباره این سامانه و روش کار آن می‌پردازیم.

پس از بررسی تعداد زیادی از ویژگی‌های محتوایی بر روی وب‌گاه‌های فارسی، در این مرحله از پژوهش از یک شناساگر جدید هرز وب فارسی برای رده‌بندی وب‌گاه‌ها استفاده نمودیم. در این سامانه پس از انجام فرآیندهای پیش‌پردازش داده مانند حذف ابربرچسب‌های HTML و حذف ایست‌واژه‌ها، از مدل جدیدی به نام BOSW برای استخراج ویژگی‌ها استفاده می‌شود. مدل BOSW در واقع مدل تغییر یافته‌ای از مدل ساده کیف کلمات است که اولین بار توسط Harris [۱۰۳] استفاده شد و بعدها نیز در زمینه‌های پردازش زبان طبیعی و بازیابی اطلاعات موارد استفاده بسیاری پیدا کرد. در مدل کیف کلمات هر سند به صورت مجموعه‌ای از nـ گرام‌های یکتا نمایش داده می‌شود. با استفاده از این روش بسیاری از اطلاعات اضافی متن مانند POS[۲]، ترتیب کلمات و سایر اطلاعات مربوط به جمله‌بندی متن حذف می‌شود. یکی از موارد استفاده مدل کیف کلمات در رده‌بندی اسناد است. بدین منظور تمام nـ گرام‌های یکتای موجود در کل مجموعه اسناد استخراج شده و به عنوان مجموعه ویژگی‌ها در نظر گرفته می‌شوند. سپس هر سند به صورت برداری از این ویژگی‌ها نمایش داده می‌شود که می‌تواند به صورت دودویی یا وزن‌دار مقداردهی شود. در روش دودویی، در صورتی که nـ گرام موردنظر در سند وجود داشته باشد درایه نظیر آن مقدار یک و در غیر این صورت مقدار صفر می‌گیرد. در روش وزن‌دار نیز از مدل‌های وزن‌دهی متفاوتی استفاده می‌شود که رایج‌ترین آن‌ها وزن‌دهی TF یا TF-IDF می‌باشد.

با توجه به آزمایشی که بر روی وب‌گاه‌های فارسی انجام دادیم، به این نتیجه رسیدیم که استفاده از مدل ساده کیف کلمات در شناسایی وب‌گاه‌های هرز فارسی کارایی خوبی ندارد. این امر به دلیل وجود خطاهایی

[۱] Persian Web Spam Detection System
[۲] Part of Speech

است که از کلمات موجود در وب‌گاه‌های معتبر ایجاد می‌شود. برخلاف صفحات وب هرز که معمولا دارای موضوع‌های مشابهی مانند موضوعات تبلیغاتی و تجاری هستند، صفحات معتبر بسته به کاربردشان دارای موضوع‌های متفاوتی می‌باشند. بنابراین، انتظار می‌رود انتخاب مجموعه کلمات رایج در صفحات هرز به عنوان بردار ویژگی، در شناسایی وب‌گاه‌های هرز تاثیر بسزایی داشته باشد. این کلمات معمولا با ترکیب خاصی در صفحات هرز ظاهر می‌شوند. در صورتی که اگر کلمات رایج در وب‌گاه‌های معتبر را نیز در بردار ویژگی در نظر بگیریم، علاوه بر افزایش هزینه محاسباتی، در مواردی نیز ممکن است باعث کاهش خطای رده‌بندی شود.

برای حل مشکلات مطرح شده، در PSD-SYS از مدل BOSW استفاده شده است. در این مدل به جای استخراج nـ گرام‌های یکتا از کل مجموعه اسناد موجود در پیکره، ابتدا مجموعه وب‌گاه‌های هرز را جدا کرده و سپس nـ گرام‌های یکتای موجود در آن‌ها را استخراج می‌کنیم و به عنوان مجموعه ویژگی‌ها در نظر می‌گیریم. سپس هر صفحه وب به صورت برداری از ویژگی‌های مشخص شده نمایش داده می شود. همان‌طور که توضیح دادیم بردار ویژگی می‌تواند با استفاده از روش‌های وزن‌دهی متفاوتی مقداردهی شود که در این سامانه پس از آزمایش روش‌های وزن‌دهی مختلف، در نهایت از روش دودویی که بهترین نتیجه را داده است استفاده می‌شود.

برای کاهش زمان پردازش و همچنین بهبود کارایی رده‌بندی نهایی، در PSD-SYS از روش‌های مختلف انتخاب ویژگی مانند χ^2-test، mutual information ، TF و TF-IDF برای ۱ ـ گرام و ۲ ـ گرام‌ها، استفاده کردیم. به ازای هر روش، ویژگی‌هایی را که امتیازشان از آستانه تعیین شده بیشتر بود به عنوان مجموعه ویژگی‌های مناسب برای رده‌بندی در نظر گرفته و بقیه ویژگی‌ها را حذف کردیم. در نهایت برای رده‌بندی وب‌گاه‌ها از الگوریتم SVM به دلیل کارایی بالاتر آن نسبت به سایر الگوریتم‌های یادگیری ماشین استفاده کردیم. نتایج آزمایش‌ها در بخش ۱.۳.۴ نشان می‌دهد که این سامانه شناساگر هرز وب فارسی نسبت به سایر روش‌های شناسایی هرز وب محتوایی دقت و فراخوانی بالاتری دارد.

۲.۳ الگوریتم‌های مبتنی بر پیوند برای شناسایی هرز وب

در این بخش از پژوهش الگوریتم‌هایی را پیشنهاد می‌دهیم که با استفاده از آن‌ها می‌توان با داشتن برچسب اعتبار (هرز) مجموعه محدودی از وب‌گاه‌ها و با بهره‌گیری از ساختار گرافی وب، امتیاز ارزشمندی (هرز بودن) سایر وب‌گاه‌ها را محاسبه کرد.

اساس کار روش‌های پیشنهادی در این بخش، الگوریتم‌های انتشار برچسب مانند TrustRank و Anti-TrustRank هستند که به صورت نیمه‌سرپرست و با داشتن امتیاز تعدادی از گره‌های درون گراف، امتیاز سایر گره‌های گراف را محاسبه می‌کنند. برای استفاده از روش‌های انتشار برچسب در حل مسائل مختلف، ابتدا باید مسئله را به صورت یک گراف مدل‌سازی کرد. بنابراین در این بخش، ابتدا به توضیح مختصری درباره نحوه مدل‌سازی گراف وب می‌پردازیم. سپس در بخش ۲.۲.۳، الگوریتمی به نام WorthyRank را معرفی می‌کنیم که با شروع از یک مجموعه بذر معتبر اولیه با امتیاز ارزشمندی مشخص، به صورت پیش‌رو امتیاز ارزشمندی آن‌ها را در کل گراف انتشار داده و در نهایت وب‌گاه‌ها را بر اساس امتیاز ارزشمندی نهایی که بدست می‌آورند رتبه‌بندی می‌کند. در این الگوریتم با استفاده از روش‌هایی مانند انتخاب بهینه گره‌های بذر، وزن‌دهی به یال‌های گراف و همچنین بسط دوره‌ای گره‌های بذر، برخی از مشکلاتی که در الگوریتم‌های انتشار برچسب پیشین وجود دارد برطرف می‌شود. در بخش ۳.۲.۳ نیز الگوریتمی به نام JunkyRank را پیشنهاد می‌دهیم که با داشتن مجموعه‌ای از وب‌گاه‌های هرز به عنوان بذر اولیه، امتیاز هرز بودن آن‌ها را هم‌زمان به صورت پس‌رو و پیش‌رو در کل گراف انتشار داده و پس از ترکیب خطی این امتیازها با امتیازهای ارزشمندی وب‌گاه‌ها که با استفاده از الگوریتم WorthyRank محاسبه شده است، در نهایت وب‌گاه‌ها را بر اساس امتیاز هرز بودنشان رتبه‌بندی می‌کند.

۱.۲.۳ مدل‌سازی گراف وب

برای استفاده از اطلاعات مربوط به ساختار پیوندی بین صفحات وب ابتدا باید صفحات وب و پیوندهای بین آن‌ها را به صورت یک گراف مدل‌سازی کرد. با فرض وجود N وب‌گاه، گراف جهت‌دار G دارای N

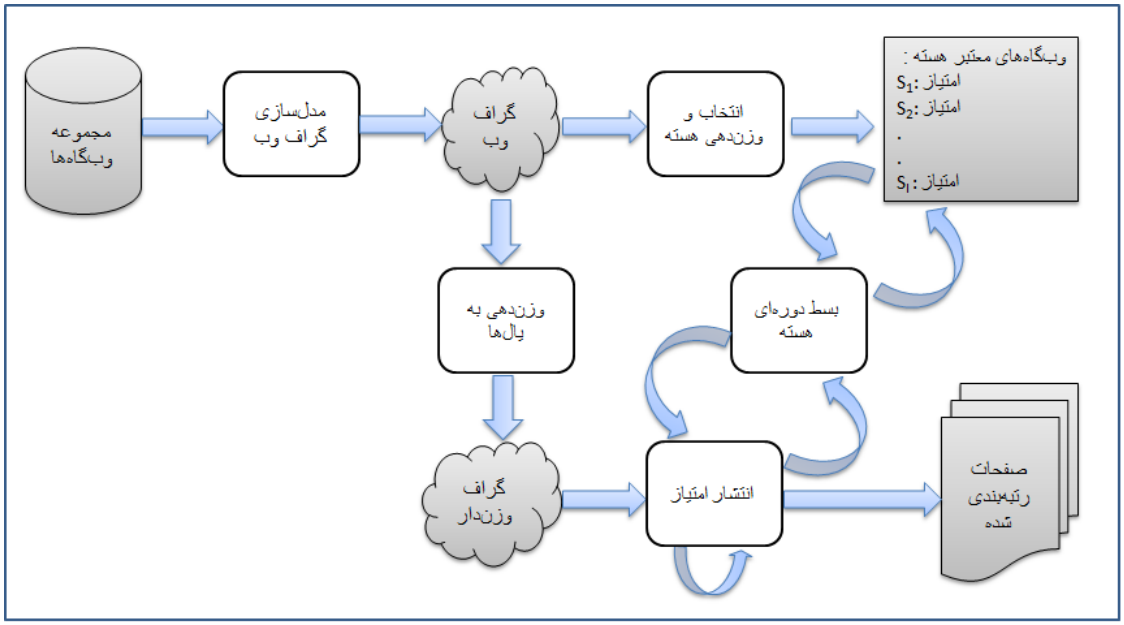
گره با مشخصه‌های ۱، ۲، ۳، ...، n است که با استفاده از یال‌های جهت‌دار به یکدیگر متصل هستند. در صورتی که A ماتریس مجاورت این گراف باشد به ازای هر زوج گره (i,j)، اگر از وب‌گاه i به یکی از صفحات وب‌گاه j پیوندی وجود داشته باشد مقدار a_{ij} برابر با یک و در غیر این صورت این مقدار برابر با صفر است. در صورت وزن‌دار بودن گراف، وزن هر یال، نشان‌دهنده تعداد پیوندهای صفحات درون وب‌گاه مبدا به صفحه یا صفحاتی در وب‌گاه مقصد می‌باشد.

۲.۲.۳ الگوریتم WorthyRank

با بررسی رفتار صفحات وب در اینترنت مشاهده می‌شود که تعداد زیادی از ارجاع‌های درون یک وب‌گاه معتبر، به صفحات معتبر دیگر می‌باشد [۱۷]. یک وب‌گاه معتبر، با توجه به این‌که دارای مطالب مفید و معتبر می‌باشد، پیوندهای درون آن نیز که مرتبط با مطالب درون صفحه هستند، پیوندهای معتبری بوده که به سایر صفحات معتبر اشاره می‌کنند. بنابراین در گراف وب، با داشتن امتیاز یک گره معتبر، می‌توان بخشی از این امتیاز را به گره‌های همسایه خروجی از آن گره انتقال داد.

با در نظر گرفتن این مهم، در این بخش الگوریتم نیمه‌سرپرستی را معرفی می‌نماییم که با داشتن تعداد اندکی وب‌گاه معتبر به عنوان بذر، و وزن‌دهی اولیه به آن‌ها، میزان امتیاز ارزشمندی آن‌ها را به صورت پیش‌رو در کل گراف انتشار می‌دهد. در روش پیش‌رو، امتیاز یک گره طبق قاعده مشخصی بین گره‌های مقصد یال‌های خروجی آن گره تقسیم می‌شود. همچنین با توجه به این‌که برچسب گره‌های بذر مشخص است، این گره‌ها همواره یک وزن مشخصی از امتیاز اولیه خود را دریافت می‌کنند. این الگوریتم به صورت دوره‌ای تا زمانی تکرار می‌شود که میزان تغییر امتیاز هر وب‌گاه در دو دور متوالی کمتر از ϵ باشد.

در نهایت تمام وب‌گاه‌ها به ترتیب بر اساس بیشترین میزان امتیاز ارزشمندی، رتبه‌بندی می‌شوند. هدف از این الگوریتم این است که وب‌گاه‌های معتبر در رتبه‌های بالاتر و وب‌گاه‌های هرز در رتبه‌های پایین‌تر قرار بگیرند. یکی از کاربردهای اصلی این الگوریتم می‌تواند در موتورهای جست‌وجو باشد. هدف موتورهای جست‌وجو این است که با استفاده از الگوریتم‌های مختلف، صفحات وب را به ترتیب بر اساس میزان اعتبار و ارتباطشان با پرس‌وجوی موردنظر به کاربر پیشنهاد دهند. در این وظیفه، هر چقدر تعداد



شکل ۱۱.۳: مراحل اجرای الگوریتم WorthyRank

صفحات هرز در میان نتایج جستجو کمتر باشد، رضایت کاربران از موتور جست‌وجو بیشتر می‌شود. شکل ۱۱.۳ مراحل اجرای الگوریتم WorthyRank را نشان می‌دهد. مطابق این شکل، این الگوریتم چند بخش اصلی دارد که در ادامه به شرح هر یک از آن‌ها می‌پردازیم.

انتخاب و وزن‌دهی گره‌های بذر

همان‌طور که در بخش ۲.۲ توضیح دادیم، روش انتخاب بذر در دقت نهایی الگوریتم‌های انتشار برچسب موثر می‌باشد. در این پژوهش برای کاهش خطای ناشی از انتخاب بذر از الگوریتم PageRank معکوس[۱] استفاده می‌نماییم. این الگوریتم مانند الگوریتم PageRank است با این تفاوت که برای محاسبه امتیاز گره‌ها، ابتدا جهت یال‌های گراف معکوس می‌شود. همان‌طور که در [۱۷] نیز توضیح داده شده است، با استفاده از این الگوریتم می‌توان گره‌هایی را که بیشترین توانایی انتشار امتیار اعتماد در کل گراف را دارند مشخص کرد. پس از محاسبه PageRank معکوس برای تمام گره‌های درون گراف، آن‌ها را به ترتیب به صورت نزولی بر اساس این امتیاز مرتب کرده و شروع به برچسب‌گذاری آن‌ها می‌نماییم. این کار را تا زمانی ادامه می‌دهیم

[۱] inverse

که مقدار گره‌های معتبر، به تعداد موردنظر رسیده باشد. با استفاده از این روش در نهایت مجموعه‌ای از L^+ وب‌گاه برچسب‌خورده معتبر داریم. با در نظر گرفتن وب‌گاه‌های بذر، بردار $\vec{S} = (s_1, s_2, ..., s_n)$ را بدست می‌آوریم که در آن s_i طبق رابطه ۴.۳ محاسبه می‌شود.

$$s_i = \begin{cases} \frac{1}{L^+}, & \text{if } i \in S^+ \\ 0, & \text{otherwise} \end{cases} \qquad (۴.۳)$$

وزن‌دهی یال‌های گراف

یکی از مشکلاتی که در گراف وب وجود دارد، وجود پیوندهای جعلی بین صفحات و وب‌گاه‌های مختلف می‌باشد. دلیل این امر هرزنویسانی هستند که برای افزایش رتبه وب‌گاه‌های هرز خود، پیوند صفحاتشان را درون وب‌گاه‌های معتبری قرار می‌دهند که به کاربران امکان نوشتن نظرات و یا مطالب دلخواه دیگر را درون بخش‌هایی از صفحات خود می‌دهند. در این پژوهش برای حل این مشکل از ضریب جاکارد[۱] برای محاسبه میزان اعتبار پیوندهای موجود در گراف وب استفاده می‌کنیم. با استفاده از این ضریب، ماتریس مجاورت A را به ماتریس وزن‌دار M تبدیل می‌کنیم که در آن $0 \leq m_{ij} \leq 1$ وزن اعتبار پیوند بین دو وب‌گاه i و j است که به صورت زیر محاسبه می‌شود:

$$m_{ij} = \frac{|N(i) \cap N(j)|}{|N(i) \cup N(j)|} \qquad (۵.۳)$$

که در آن $N(i)$ مجموعه تمام همسایه‌های گره i است. با توجه به رابطه ۵.۳ امتیازی که با استفاده از این روش به پیوند بین دو وب‌گاه داده می‌شود بر اساس میزان اشتراک همسایه‌های دو وب‌گاه است. پیوند بین دو وب‌گاه نمایانگر وجود یک ارتباط بین دو وب‌گاه است. در نتیجه در صورتی که این ارتباط واقعی باشد احتمال وجود اشتراک بین همسایه‌های این دو وب‌گاه نیز زیاد است. بنابراین با استفاده از این روش یال‌های گراف اعتبارسنجی شده و یال‌های مربوط به پیوندهای جعلی که توسط هرزنویسان درون صفحات مختلف

[۱] Jaccard coeficient

قرار داده شده است وزن کمی می‌گیرند.

انتشار امتیاز ارزشمندی

پس از مشخص کردن مجموعه گره‌های بذر و وزن‌دهی آن‌ها، در هر تکرار t، امتیاز ارزشمندی گره i به صورت زیر محاسبه می‌شود:

$$r_i(t) = \alpha \sum_{j=1}^{n} m_{ji} r_j(t-1) + (1-\alpha) * s_i \qquad (۶.۳)$$

در این رابطه α یک عامل میرایی[۱] است که مشخص می‌کند چند درصد از امتیاز ارزشمندی هر صفحه از امتیاز سایر صفحاتی که به آن ارجاع داده‌اند و چند درصد آن، از مقدار وزن مشخصی که به صفحات بذر داده می‌شود تامین شود. همچنین مقدار s_i همان مقدار مربوط به وزن اولیه صفحات بذر است که با استفاده از رابطه ۴.۳ محاسبه شده است.

بسط دوره‌ای گره‌های بذر

همان‌طور که در بخش ۲.۲ توضیح دادیم، در پژوهش‌هایی که بر روی تاثیر اندازه مجموعه گره‌های بذر بر روی نتیجه نهایی الگوریتم‌های انتشار امتیاز مانند TrustRank [۱۷] انجام شده است، نشان داده‌اند که کم بودن تعداد بذرهای اولیه می‌تواند باعث کاهش دقت رتبه‌بندی نهایی شود. از طرف دیگر برچسب‌گذاری وب‌گاه‌ها به طور دستی کاری بسیار زمان‌بر و پرهزینه می‌باشد. بنابراین در الگوریتم WorthyRank از روش جدیدی به نام بسط دوره‌ای بذر استفاده می‌کنیم که با داشتن تعداد کمی بذر اولیه، در هر مرحله تعداد وب‌گاه‌های بذر را به طور خودکار افزایش می‌دهد. در این روش ابتدا تعداد محدودی از گره‌های معتبر را به عنوان گره‌های بذر انتخاب می‌نماییم. سپس در هر بار تکرار الگوریتم، امتیاز اعتماد گره‌ها را محاسبه کرده و آن‌ها را به صورت نزولی مرتب می‌نماییم. سپس k گره با بیشترین امتیاز ارزشمندی را به عنوان بذرهای

[۱] damping factor

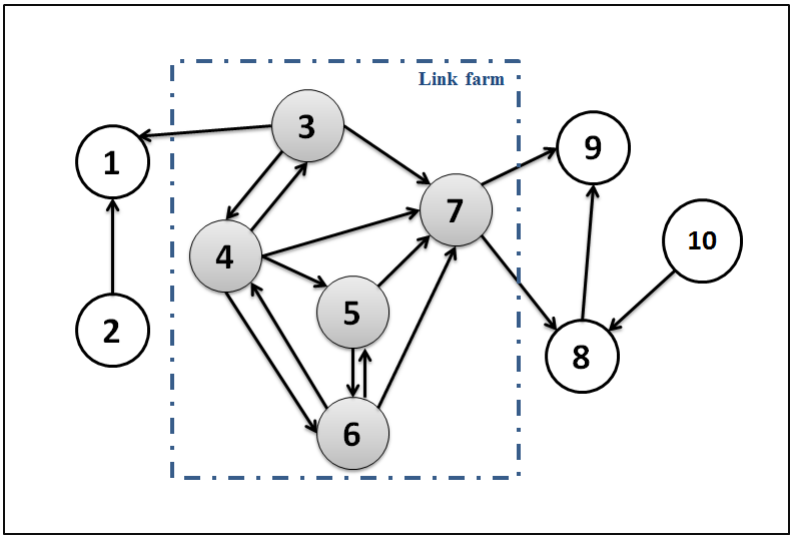
جدید به مجموعه بذر اولیه اضافه می‌کنیم. این کار را تا جایی ادامه می‌دهیم که تعداد گره‌های بذر به f درصد تعداد کل گره‌های گراف برسد.

۳.۲.۳ الگوریتم JunkyRank

الگوریتم‌های انتشار برچسب هرز مانند Anti-TrustRank [۶۰]، DRank [۷۲] و BRank [۷۳] بر اساس این فرض عمل می‌کنند که صفحاتی که به یک صفحه هرز ارجاع می‌دهند به احتمال زیاد جزء صفحات هرز می‌باشند. بر این اساس، در این الگوریتم‌ها امتیاز هرز بودن هر صفحه در مسیر عکس جهت یال‌های گراف به تمام صفحاتی که به آن صفحه ارجاع داده‌اند انتشار داده می‌شود. مشکلی که در این دسته از الگوریتم‌ها وجود دارد این است که هیچ احتمال هرز بودنی برای صفحاتی که یک صفحه هرز به آن‌ها ارجاع داده است در نظر گرفته نمی‌شود. این امر در حالی است که تعداد زیادی از صفحات هرز برای افزایش رتبه‌شان، به یکدیگر ارجاع می‌دهند. این‌گونه از صفحات که معمولا بخشی از یک دهکده‌پیوندی هستند، با تبادل پیوند با یکدیگر، رتبه وب‌گاه‌های خود را نسبت به سایر وب‌گاه‌ها افزایش می‌دهند.

برای مثال در شکل ۱۲.۳ که بخشی از یک گراف وب را نشان می‌دهد، مجموعه گره‌های ۳، ۴، ۵، ۶ و ۷، با یکدیگر یک دهکده پیوندی را تشکیل می‌دهند. در این گراف، گره‌های سفید مربوط به وب‌گاه‌های معتبر و گره‌های خاکستری نشان‌گر وب‌گاه‌های هرز می‌باشند. با انتخاب گره‌های ۴ و ۶ به عنوان بذر هرز، در صورتی که برای محاسبه امتیاز هرز سایر گره‌ها، فقط از انتشار امتیاز در مسیر یال‌های گراف استفاده شود، گره ۷ هیچ امتیاز هرزی دریافت نمی‌کند. با توجه به این‌که در الگویتم JunkyRank امتیاز هرز گره‌های بذر به صورت پیش‌رو نیز در گراف انتشار داده می‌شود، گره ۷ در رتبه‌بندی بر اساس امتیاز هرز توسط این الگوریتم، در رتبه‌های بالاتر قرار می‌گیرد.

برای حل این مشکل الگوریتمی به نام JunkyRank را معرفی می‌نماییم که با داشتن مجموعه‌ای از وب‌گاه‌های هرز به عنوان بذر اولیه، امتیاز هرز بودن آن‌ها را هم‌زمان به دو صورت پیش‌رو و پس‌رو در گراف وب انتشار می‌دهد. این الگوریتم دو بخش اصلی دارد که در ادامه به شرح هر یک از آن‌ها می‌پردازیم.



شکل ۱۲.۳: بخشی از گراف وب که دارای یک دهکده پیوندی می‌باشد.

انتخاب و وزن‌دهی گره‌های بذر

در این بخش برای انتخاب گره‌های بذر، از الگوریتم HITS [۱۶] استفاده می‌نماییم. در این الگوریتم برای هر گره در گراف وب دو امتیاز محاسبه می‌شود. امتیاز hub برای یک گره، مشخص می‌کند که یک گره چه میزان پیوند خروجی به گره‌های با authority بالا دارد. همچنین امتیاز authoruty نیز برای یک گره، نشان‌گر میزان ارجاعات گره‌های مختلف با امتیاز hub بالا به آن گره می‌باشد. با توجه به این‌که امتیاز گره‌های بذر در الگوریتم JunkyRank، هم به صورت هم‌جهت با یال‌های گراف و هم در خلاف جهت یال‌های گراف انتشار داده می‌شود، گره‌هایی به عنوان بذر مناسب هستند که مقدار hub و یا authority بالایی داشته باشند.

برای انتخاب گره‌های بذر در این الگوریتم، ابتدا الگوریتم HITS را بر روی گراف وب اجرا می‌کنیم، سپس برای انتخاب l گره به عنوان بذر، $\frac{l}{2}$ آن را از گره‌های با بیشترین امتیاز hub و $\frac{l}{2}$ آن را از گره‌های با بیشترین امتیاز authorithy انتخاب می‌نماییم. دلیل این امر، انتخاب گره‌هایی به عنوان بذر است که قابلیت بالایی را در انتشار امتیاز به صورت پیش‌رو و پس‌رو در کل گراف دارند. با استفاده از این روش در نهایت مجموعه‌ای از l^- وب‌گاه‌ها برچسب‌خورده هرز داریم. با توجه به این مجموعه وب‌گاه‌های بذر، بردار

$\vec{S'} = (s'_1, s'_2, ..., s'_n)$ را بدست می‌آوریم، به طوری‌که در آن، مقدار s'_i طبق رابطه ۷.۳ محاسبه می‌شود.

$$s'_i = \begin{cases} \frac{1}{L^-}, & \text{if } i \in S^- \\ 0, & \text{otherwise} \end{cases} \qquad (۷.۳)$$

انتشار امتیاز هرز بودن

همان‌طور که توضیح دادیم، مزیت الگوریتم JunkyRank نسبت به روش‌های پیشین این است که علاوه بر انتشار امتیاز هرز بودن در خلاف جهت یال‌ها، این امتیاز را در جهت یال‌های گراف وب نیز انتشار می‌دهد. نکته‌ای که باید در نظر داشت این است که احتمال هرز بودن یک وب‌گاه، در صورتی که به یک وب‌گاه هرز ارجاع داده باشد، بیشتر از حالتی است که توسط یک وب‌گاه هرز ارجاع داده شده باشد. دلیل این امر این است که وب‌گاه‌های هرز در خیلی از موارد برای افزایش رتبه خود، به وب‌گاه‌های معتبر ارجاع می‌دهند. بنابراین برای کاهش اثر منفی وجود این نوع از پیوندهای جعلی در گراف وب، در این الگوریتم از ضریبی به نام β استفاده می‌شود که به یال‌های ورودی به یک صفحه و یال‌های خروجی از آن وزن متفاوتی می‌دهد.

رابطه ۸.۳ نحوه محاسبه امتیاز هرز بودن وب‌گاه‌ها را در الگوریتم JunkyRank نشان می‌دهد. با استفاده از ضریب β در این رابطه، می‌توان بخش کمی از امتیاز هرز یک گره را در جهت یال‌های گراف و وزن بیشتری از آن را در خلاف جهت یال‌های آن گره انتشار داد.

$$r_i(t) = \alpha \left(\beta \sum_{j=1}^{n} a_{ij} r_j(t-1) + (1-\beta) \sum_{j=1}^{n} a_{ji} r_j(t-1) \right) + (1-\alpha) * s'_i \qquad (۸.۳)$$

پس از انتشار امتیاز هرز بودن گره‌های بذر در گراف و رسیدن به حالت همگرایی[۱]، برای هر وب‌گاه یک امتیاز هرز بودن بدست می‌آید. اما استفاده از این امتیاز برای رده‌بندی وب‌گاه‌ها به تنهایی کافی نمی‌باشد. دلیل این امر وجود وب‌گاه‌های هرزی است که برای بالا بردن رتبه خود، پیوند تعداد زیادی از وب‌گاه‌های

[۱] convergence

معتبر را درون صفحاتشان می‌گذارند.

در صورتی که برای رتبه‌بندی نهایی وب‌گاه‌ها، فقط از امتیاز بدست آمده از الگوریتم JunkyRank استفاده کنیم، امتیاز هرز بودن این‌گونه از وب‌گاه‌های معتبر که مورد هجوم وب‌گاه‌های هرز قرار گرفته‌اند بالا می‌رود. در صورتی که این امتیاز صحیح نمی‌باشد و برای مشخص کردن این امر کافی است که امتیاز ارزشمندی این‌گونه وب‌گاه‌های معتبر نیز محاسبه شود. در نتیجه، برای رتبه‌بندی این وب‌گاه‌ها بر اساس احتمال هرز بودن آن‌ها، ابتدا ترکیب خطی امتیاز هرز بودن وب‌گاه‌ها در الگوریتم JunkyRank را با امتیاز ارزشمندی آن‌ها در الگوریتم WorthyRank با استفاده از رابطه ۹.۳ بدست آورده و در نهایت وب‌گاه‌ها را بر اساس امتیاز حاصله از این ترکیب خطی رتبه‌بندی می‌کنیم.

$$JunkyRank_{\gamma} = \gamma * JunkyRank - (1-\gamma) * WorthyRank \qquad (۹.۳)$$

برای مثال در شکل ۱۲.۳، در صورتی که برای شناسایی وب‌گاه‌های هرز، فقط امتیاز هرز بودن گره‌ها را در نظر بگیریم، گره‌های معتبر ۱ و ۸ و ۹ نیز به عنوان صفحات هرز شناسایی می‌شوند. در حالی‌که با انتخاب دو گره ۲ و ۱۰ به عنوان بذر معتبر و استفاده از ترکیب خطی امتیاز هرز با امتیاز اعتمادی که گره‌ها با استفاده از الگوریتم WorthyRank کسب می‌کنند، می‌توان کیفیت رتبه‌بندی را بهبود داد.

۴.۲.۳ اثبات همگرایی

برای اثبات همگرایی یک الگوریتم، ابتدا لازم است که شرط یا شرایطی را برای همگرایی آن تعریف نماییم. در الگوریتم‌های مبتنی بر پیوندی که در این پژوهش معرفی کرده‌ایم شرطی که برای همگرایی در نظر می‌گیریم به صورت رابطه ۱۰.۳ می‌باشد.

$$\forall i : |r_i(t) - r_i(t-1)| \leq \epsilon \Longleftrightarrow \lim_{t\to\infty} \frac{r_i(t)}{r_i(t-1)} = 1 \qquad (۱۰.۳)$$

طبق این رابطه، اجرای الگوریتم تا زمانی ادامه می‌یابد که مقدار اختلاف امتیاز هر گره، در دو دور متوالی از اجرای الگوریتم، کمتر از ϵ باشد.

تمام الگوریتم‌هایی که در این بخش معرفی شده‌اند دارای یک پایه مشترک با یکدیگر هستند و آن میزان احتمالی است که وزن انتشار امتیاز از گره i به گره j را مشخص می‌کند و این‌که این وزن برخلاف بسیاری از روش‌های پیشین هنجارسازی[۱] نمی‌شود. سایر سیاست‌های به‌کار گرفته شده در الگوریتم‌ها، مانند وزن‌دهی به گره‌های بذر، بسط مجموعه گره‌های بذر، وجود عامل میرایی و ترکیب خطی امتیاز ارزشمندی و غیر ارزشمندی تاثیری در همگرایی الگوریتم‌های انتشار برچسب ندارند. دلیل این امر این است که تمام موارد ذکر شده، به صورت یک پارامتر ثابت در روابط هر یک از الگوریتم‌ها ظاهر می‌شوند. تنها تاثیری که این پارامترها می‌توانند در الگوریتم‌های مربوطه داشته باشند، تاثیر آن‌ها در کارایی نهایی الگوریتم و همچنین زمان رسیدن به شرط همگرایی می‌باشد. با در نظر گرفتن این مهم، برای اثبات همگرایی الگوریتم‌های انتشار برچسب که در این بخش معرفی کرده‌ایم کافی است ثابت کنیم معادله ساده شده ۱۱.۳ شرط همگرایی ۱۰.۳ را ارضا می‌کند.

$$r_i(t) = \sum_{j=1}^{n} m_{ji} r_j(t-1) \qquad (۱۱.۳)$$

با تعریف بردار $R(t) = (r_1(t), r_2(t), ..., r_n(t))$ به عنوان بردار امتیاز صفحات در زمان t که در آن $r_i(t)$ امتیاز وب‌گاه iام در زمان t است، می‌توان معادله ۱۱.۳ را به صورت ماتریسی به صورت معادله ۱۲.۳ نوشت.

$$R(t) = R(t-1)M \Longrightarrow R(t) = R(0)M^t \qquad (۱۲.۳)$$

[۱] normalization

طبق قضیه ۸-۵-۱ از کتاب [۱۰۴] برای یک ماتریس نامنفی و اصلی[۲] A داریم:

$$\lim_{m\to\infty}(\rho(A)^{-1}A)^m = L > 0 \tag{۱۳.۳}$$

به طوری‌که $L = XY^T$، $AX = \rho(A)X$، $A^TY = \rho(A)Y$ و $X, Y > 0$ است.

همچنین طبق قضیه ۸-۵-۲ در [۱۰۴]، یک ماتریس نامنفی A یک ماتریس اصلی است اگر و تنها اگر $m \geq 1$ وجود داشته باشد که به ازای آن داشته باشیم $A^m > 0$. با توجه به این‌که درایه‌های ماتریس مجاورت وزن‌دار M در الگوریتم‌های معرفی شده، هر کدام مقداری بین ۰ و ۱ دارند، این ماتریس یک ماتریس نامنفی می‌باشد. همچنین با توجه به این‌که این ماتریس مربوط به گراف وب است و گراف وب یک گراف قویا متصل[۱] است بنابراین به ازای هر زوج گره i و j، حداقل یک مسیر از i به j وجود دارد. به عبارت دیگر $t \geq 1$ وجود دارد به طوری‌که به ازای آن $M^t > 0$ است. در نتیجه M^t یک ماتریس اصلی است. بنابراین با توجه به معادله ۱۳.۳ داریم:

$$\lim_{t\to\infty}(\lambda_1^{-1}M)^t = V_1V_2^T \Longrightarrow M^t = \lambda_1^tV_1V_2^T \tag{۱۴.۳}$$

با جایگذاری مقدار M^t در رابطه ۱۲.۳ داریم:

$$R(t) = R(0)\lambda_1^tV_1V_2^T \Longrightarrow r_i(t) = \lambda_1^t\sum_{k=1}^{n} v_{1k}v_{2i}r_k(0) \tag{۱۵.۳}$$

در نتیجه برای شرط همگرایی ۱۰.۳ داریم:

$$\lim_{t\to\infty}\frac{r_i(t)}{r_i(t-1)} = \frac{\lambda_1^t\sum_{k=1}^{n} v_{1k}v_{2i}r_k(0)}{\lambda_1^{t-1}\sum_{k=1}^{n} v_{1i}v_{2k}s_k(0)} = \lambda_1 \tag{۱۶.۳}$$

[۲] primitive
[۱] strongly connected

با توجه به رابطه ۱۶.۳ مشاهده می‌کنیم در صورتی که امتیاز تمام گره‌ها پس از هر دور اجرای الگوریتم به‌هنجار[۲] نشود، شرط همگرایی برقرار نمی‌شود. بنابراین پس از اعمال هنجارسازی امتیازها در رابطه ۱۶.۳ داریم:

$$\lim_{t\to\infty}\frac{r_i(t)}{r_i(t-1)}=\frac{\left(\frac{\lambda_1^t\sum_{k=1}^{n}v_{1k}v_{2i}r_k(0)}{\lambda_1^t\sum_{i=1}^{n}\sum_{k=1}^{n}v_{1k}v_{2i}r_k(0)}\right)}{\left(\frac{\lambda_1^{t-1}\sum_{k=1}^{n}v_{1i}v_{2k}s_k(0)}{\lambda_1^{t-1}\sum_{i=1}^{n}\sum_{k=1}^{n}v_{1k}v_{2i}r_k(0)}\right)}=1 \qquad (۱۷.۳)$$

همان‌طور که در رابطه ۱۷.۳ مشاهده می‌کنیم شرط همگرایی ۱۰.۳ برقرار شده است.

۳.۳ روش ترکیبی محتوایی و پیوندی برای شناسایی هرز وب

در این بخش، برای بهبود کارایی رتبه‌بندی وب‌گاه‌ها بر اساس الگوریتم انتشار برچسب، الگوریتمی به نام CLCRank[۱] را پیشنهاد می‌دهیم که برای رتبه‌بندی وب‌گاه‌ها علاوه بر استفاده از اطلاعات پیوندی بین آن‌ها، از اطلاعات مربوط به محتوای وب‌گاه‌ها نیز استفاده می‌نماید.

همان‌طور که در بخش‌های قبل توضیح دادیم، هرزنویسان برای افزایش رتبه صفحات خود، علاوه بر استفاده از روش‌های هرزنویسی محتوایی، از روش‌های پیوندی نیز استفاده می‌کنند. آن‌ها با استفاده از ترکیب روش‌های مختلف سعی می‌کنند بسیاری از الگوریتم‌های شناسایی هرز وب را فریب دهند. بنابراین برای شناسایی و مقابله با این نوع از صفحات، به روش‌هایی نیاز است که از انواع مختلفی از اطلاعات مربوط به یک صفحه، برای بررسی هرز یا معتبر بودن آن استفاده می‌کنند. برای مثال ممکن است یک وب‌گاه، نسبت به سایر وب‌گاه‌ها امتیاز پیوندی بالایی را بدست آورد، اما محتوای درون آن وب‌گاه شامل مطالب غیرمفید و هرز باشد. این دسته از وب‌گاه‌ها، وب‌گاه‌هایی هستند که پیوند صفحات خود را درون صفحات معتبر زیادی قرار می‌دهند. همچنین در گراف‌های وب، معمولا تعدادی گره وجود دارند که با سایر بخش‌های گراف

[۲] normalized
[۱] Combined Link-based and Content-based Ranking

ارتباط چندانی ندارند. در صورتی که برای بررسی میزان اعتبار این نوع از صفحات فقط از روش‌های پیوندی استفاده شود، کیفیت رتبه‌بندی کاهش می‌یابد. از طرف دیگر، روش‌های شناسایی هرز وب با استفاده از محتوای صفحات، به تنهایی قادر به شناسایی تمام وب‌گاه‌های هرز نمی‌باشند. برای مثال، صفحات هرزی وجود دارند که به ظاهر دارای محتوای مفید و مطالب معتبر می‌باشند. اما با بررسی ویژگی‌های پیوندی این صفحات در گراف وب، درمی‌یابیم که هیچ صفحه معتبری به این صفحات ارجاع نداده است. این امر نشان می‌دهد که مطالب مفید درون این صفحات وب، صرفا یک رونوشت از مطالب وب‌گاه‌های معتبر می‌باشد. برای شناسایی این نوع از وب‌گاه‌ها می‌توان از روش‌های ترکیبی مانند CLCRank که در این پژوهش ارائه می‌شود استفاده نمود.

در الگوریتم CLCRank ابتدا با استفاده از مجموعه‌ای از وب‌گاه‌های برچسب‌خورده و آموزش آن‌ها، احتمال ارزشمندی سایر وب‌گاه‌ها را محاسبه می‌نماییم. سپس برای انتشار امتیاز ارزشمندی یا هرز بودن وب‌گاه‌های بذر در کل گراف، میزان انتشار امتیاز از طریق هر یال، با استفاده از احتمال اعتبار یا هرز محتوایی گره‌های مبدا و مقصد تعیین می‌شود. لازم به ذکر است که برای اجرای این الگوریتم نیز مشابه روشی که در بخش ۱.۲.۳ برای الگوریتم‌های پیوندی توضیح دادیم، ابتدا با استفاده از پیوند بین وب‌گاه‌ها، گراف وب را مدل‌سازی می‌نماییم. شرح کامل مراحل اصلی این الگوریتم در ادامه ارائه شده است.

۱.۳.۳ انتخاب گره‌های بذر و وزن‌دهی محتوایی

در مرحله اول، برای اجرای الگوریتم CLCRank لازم است برای هر وب‌گاه، یک امتیاز ارزشمندی محتوایی محاسبه شود. بدین منظور می‌توان از هر یک از روش‌های رده‌بندی مبتنی بر محتوا استفاده نمود. نکته‌ای که در این مرحله باید در نظر داشت، انتخاب تعدادی از وب‌گاه‌ها به عنوان داده‌های آموزش می‌باشد. همان‌طور که می‌دانیم، الگوریتم‌های انتشار برچسب، الگوریتم‌های نیمه‌سرپرستی هستند که با داشتن برچسب تعداد محدودی از گره‌های گراف، برچسب سایر گره‌ها را پیش‌بینی می‌کنند. در این روش‌ها، تعداد داده‌های برچسب‌خورده به نسبت تعداد کل داده‌ها کم است. بنابراین انتخاب وب‌گاه‌های بذر در الگوریتم‌های

پیوندی به عنوان داده‌های آموزش، برای رده‌بندی محتوایی وب‌گاه‌ها که بیشتر آن‌ها به صورت باسرپرست[۱] هستند، کافی نمی‌باشد.

بنابراین در الگوریتم CLCRank، علاوه بر استفاده از وب‌گاه‌های بذر به عنوان داده‌های آموزش، از وب‌گاه‌هایی که دارای ارتباط کمتری با سایر بخش‌های گراف هستند نیز استفاده می‌نماییم. با توجه به این‌که این نوع از وب‌گاه‌ها اطلاعات پیوندی زیادی ندارند، می‌توان در ازای آن، از اطلاعات محتوایی آن‌ها استفاده نمود. برای پیدا کردن این وب‌گاه‌ها، ابتدا گراف جهت‌دار وب را به یک گراف بدون جهت تبدیل می‌کنیم و مقدار PageRank آن را محاسبه می‌نماییم. سپس k گره با کمترین امتیاز را به عنوان داده‌های آموزش انتخاب می‌کنیم.

علاوه بر انتخاب گره‌های با کمترین مقدار PageRank به عنوان مجموعه وب‌گاه‌های آموزش، تعدادی وب‌گاه بذر معتبر S و هرز S' را نیز با روش‌های پیوندی انتخاب بذر مانند روشی که در بخش ۲.۲.۳ و ۳.۲.۳ توضیح دادیم انتخاب کرده و به مجموعه داده‌های آموزش اضافه می‌نماییم. سپس ویژگی‌های محتوایی این مجموعه وب‌گاه‌ها را استخراج کرده و با آموزش آن‌ها یک رده‌بند محتوایی برای وب‌گاه‌ها می‌سازیم. با استفاده از این رده‌بند محتوایی، در نهایت برای هر وب‌گاه i در گراف وب، یک احتمال ارزشمندی P_i و در نتیجه احتمال هرز بودن $P'_i = 1 - P_i$ محاسبه می‌شود.

۲.۳.۳ انتشار امتیاز

پس از مشخص کردن احتمال ارزشمندی محتوایی وب‌گاه‌ها و با داشتن مجموعه‌ای از وب‌گاه‌های معتبر S و هرز S' به عنوان بذر، از ترکیب این احتمال‌های محتوایی با الگوریتم‌های پیوندی فصل ۲.۳ استفاده می‌نماییم.

یکی از مشکلاتی که در انتشار امتیاز در گراف وب وجود دارد، وجود یال‌های جعلی در این گراف است که باعث می‌شود در مواردی خاص، امتیاز هرز بودن یک وب‌گاه هرز از طریق این پیوندهای جعلی به اشتباه به یک وب‌گاه معتبر انتشار داده شود. همچنین در مواردی که یک وب‌گاه هرز، پیوند صفحات خود

[۱] supervised

را درون صفحات معتبر قرار می‌دهد، در الگوریتم‌های پیوندی می‌تواند بخشی از امتیاز اعتماد آن صفحه را بدست آورد. با در نظر گرفتن احتمال محتوایی وب‌گاه‌ها، می‌توان میزان انتشار امتیاز گره‌های گراف وب را بر اساس امتیاز محتوایی آن‌ها تعیین نمود. با استفاده از این روش، تاثیر منفی ناشی از وجود پیوندهای جعلی در گراف وب کاهش می‌یابد.

بدین منظور در الگوریتم CLCRank ابتدا احتمال اعتبار یال‌های گراف را با استفاده از احتمال ارزشمندی محتوایی وب‌گاه‌های مبدا و مقصد محاسبه کرده و گراف وب را تبدیل به یک گراف وزن‌دار می‌نماییم. سپس امتیاز اولیه وب‌گاه‌های بذر، طبق یک قاعده مشخص در این گراف وزن‌دار انتشار داده می‌شود.

در این الگوریتم برای محاسبه امتیاز اعتبار وب‌گاه‌ها و رتبه‌بندی آن‌ها بر اساس این امتیاز، از ترکیب الگوریتم WorthyRank با احتمال ارزشمندی محتوایی وب‌گاه‌ها استفاده می‌شود. بنابراین برای ترکیب این احتمال، به‌جای استفاده از ماتریس مجاورت M در رابطه ۶.۳ مربوط به الگوریتم WorthyRank، از ماتریس وزن‌دار U استفاده می‌کنیم که در آن به ازای هر زوج گره i و j داریم:

$$u_{ij} = \frac{p_i + p_j}{2} * m_{ij} \qquad (۱۸.۳)$$

با توجه به این رابطه، وزن اعتبار هر یال u_{ij}، بر اساس میزان اعتبار محتوایی گره مبدا (i) و گره مقصد (j) محاسبه می‌شود. بنابراین با استفاده از این رابطه، یک وب‌گاه هرز نمی‌تواند وزن زیادی از امتیاز اعتماد وب‌گاه‌های معتبر را بدست آورد.

همچنین برای ترکیب احتمال ارزشمندی محتوایی با الگوریتم JunkyRank، به جای استفاده از ماتریس مجاورت A در رابطه ۸.۳ مربوط به الگوریتم JunkyRank، از ماتریس وزن‌دار Y استفاده می‌نماییم که به ازای هر زوج گره i و j، درایه y_{ij} با استفاده از رابطه ۱۹.۳ محاسبه می‌شود.

$$y_{ij} = \frac{p'_i + p'_j}{2} * a_{ij} \qquad (۱۹.۳)$$

با استفاده از ماتریس Y میزان انتشار امتیاز هرز بودن از گره i به گره j با توجه به وزن هرز بودن محتوایی دو گره مبدا و مقصد کنترل می‌شود. این امر موجب می‌شود که از انتشار امتیاز هرز بودن یک وب‌گاه هرز به وب‌گاه معتبر تا حد زیادی جلوگیری شود.

با توجه به این‌که شرایط اجرای الگوریتم CLCRank مانند الگوریتم‌های پیوندی ارائه شده در فصل ۲.۳ است، اثبات همگرایی این الگوریتم نیز مشابه آن‌چه که در بخش ۴.۲.۳ بیان شده است می‌باشد. نتایج بخش آزمایش‌ها نشان می‌دهد که ترکیب روش محتوایی با روش پیوندی، دقت رده‌بندی وب‌گاه‌ها را افزایش می‌دهد.

فصل ۴

ارزیابی

در این فصل به ارزیابی روش‌های ارائه شده در شناسایی هرز وب می‌پردازیم. بدین منظور، ابتدا به معرفی مجموعه‌های داده‌ای مورد استفاده در این پژوهش پرداخته و سپس، معیارهایی را که برای ارزیابی هر یک از روش‌های معرفی شده در این پژوهش استفاده شده‌اند معرفی می‌نماییم. در انتها نیز نتایج آزمایش‌هایی را که در راستای انجام این پژوهش صورت گرفته است ارائه داده و با سایر روش‌های پیشین مربوطه مقایسه می‌نماییم.

۱.۴ مجموعه‌های داده‌ای

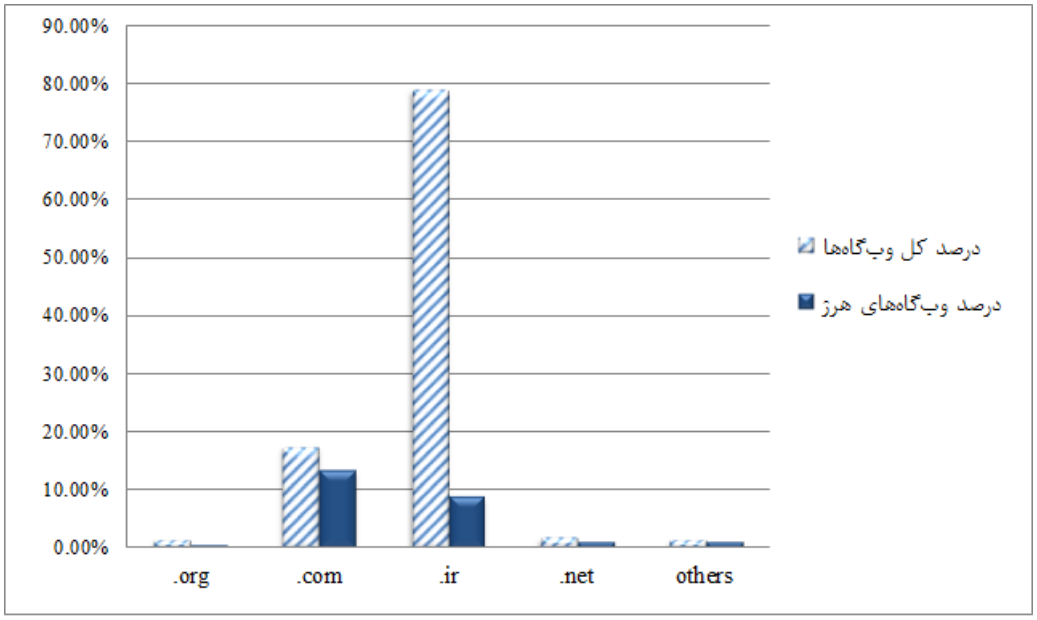
در این پژوهش، برای ارزیابی روش‌های معرفی شده در وظیفه شناسایی هرز وب، از یک مجموعه داده‌ای فارسی و دو مجموعه داده‌ای انگلیسی استفاده شده است. با توجه به خاصیت وابستگی به زبان روش‌های مبتنی بر محتوا و این‌که تمرکز این پژوهش بر روی زبان فارسی می‌باشد، برای ارزیابی این روش‌ها از مجموعه داده‌ای PersianWebSpam-2013 که شامل وب‌گاه‌های فارسی می‌باشد استفاده شده است. به دلیل محدود بودن این مجموعه داده‌ای به وب‌گاه‌های فارسی و همچنین کم بودن تعداد وب‌گاه‌های آن در مقایسه با کل وب، امکان ایجاد گراف مربوط به پیوندهای بین این وب‌گاه‌ها وجود نداشت. بنابراین برای ارزیابی روش‌های

مبتنی بر پیوند و روش ترکیبی محتوایی و پیوندی، از دو مجموعه داده‌ای استاندارد WebSpamChallengeII-CorpusI و WEBSPAM-UK2007، که دارای گراف مربوط به پیوندهای بین وب‌گاه‌ها هستند، استفاده شده است. همچنین به دلیل محدودیتی که در دسترسی به اطلاعات محتوایی هر یک از این دو مجموعه داده‌ای استاندارد وجود دارد، امکان ارزیابی روش‌های محتوایی ارائه شده در این پژوهش بر روی این دو مجموعه داده‌ای استاندارد وجود نداشت. خصوصیات هر یک از این مجموعه‌های داده‌ای در ادامه شرح داده می‌شود.

۱.۱.۴ مجموعه داده‌ای روش‌های مبتنی بر محتوا

با توجه به این‌که هدف از این بخش از پژوهش، بررسی میزان تاثیر ویژگی‌های محتوایی بر روی شناسایی وب‌گاه‌های هرز فارسی و ارائه ویژگی‌ها و روش‌های محتوایی جدید برای بهبود این وظیفه بوده است، از مجموعه داده‌ای PersianWebSpam-2013 که شامل وب‌گاه‌های فارسی می‌باشد استفاده شده است. این مجموعه داده‌ای، که به منظور انجام این پژوهش ایجاد شده است، شامل ۱۰۵۰ وب‌گاه معتبر و ۳۰۰ وب‌گاه هرز می‌باشد که در فاصله می ۲۰۱۳ تا آگوست ۲۰۱۳ خزش و برچسب‌گذاری شده‌اند.

برای داشتن یک نمونه مناسب از وب‌گاه‌های هرز فارسی، تمامی وب‌گاه‌های موجود در این مجموعه داده‌ای به صورت تصادفی و بدون اعمال محدودیت در دامنه آن‌ها جمع‌آوری شده‌اند. شکل ۱.۴ نشان‌گر نحوه توزیع این مجموعه وب‌گاه‌ها بر روی دامنه‌های مختلف و همچنین میزان پراکندگی وب‌گاه‌های هرز در هر یک از دامنه‌ها می‌باشد. با توجه به شکل ۱.۴ مشاهده می‌نماییم که حدود ۸۰ درصد از وب‌گاه‌های فارسی در این مجموعه داده‌ای، مربوط به دامنه ir. بوده، که در این میان ۸/۷ درصد آن را وب‌گاه‌های هرز تشکیل می‌دهند. پس از دامنه ir.، دامنه‌ای که بیشترین تعداد وب‌گاه‌ها را شامل می‌شود، دامنه com. با حدود ۱۷۷ وب‌گاه هرز می‌باشد. این آمار نشان می‌دهد که دامنه com. که مربوط به وب‌گاه‌های تجاری می‌باشد، جایگاه مناسبی برای بسیاری از هرزنویسان است. این دسته از هرزنویسان، با استفاده از روش‌های هرزنویسی مختلف، سعی دارند رتبه وب‌گاه‌های خود را در میان نتایج موتورهای جست‌وجو بالا ببرند، تا بتوانند محصولات خود را تبلیغ کنند و سبب جذب مشتری‌های بیشتر و در نهایت، فروش و سود تجاری



شکل ۱.۴: توزیع وب‌گاه‌های فارسی بر روی دامنه‌های مختلف در مجموعه داده‌ای PersianWebSpam-2013

بیشتر شوند.

۲.۱.۴ مجموعه داده‌ای روش‌های مبتنی بر پیوند و روش ترکیبی

به منظور ارزیابی روش‌های مبتنی بر پیوند و همچنین روش ترکیبی معرفی شده در این پژوهش، از دو مجموعه داده‌ای استاندارد WebSpamChallengeII-CorpusI [۱۰۵] و WEBSPAM-UK2007 [۱۰۶] استفاده شده است. خصوصیات هر یک از این مجموعه‌های داده‌ای به شرح زیر می‌باشد:

- مجموعه داده‌ای WebSpamChallengeII-CorpusI: این مجموعه‌داده‌ای که مربوط به وب‌گاه‌های خزش شده در سال ۲۰۰۶ می‌باشد، در مسابقه‌ای به نام Web Spam Challenge II[۱] در سال ۲۰۰۷ مورد استفاده قرار گرفته است. این مجموعه داده‌ای شامل ۹۰۷۲ وب‌گاه است که در این میان ۱۹۳۴ وب‌گاه (۲۱٪)، برچسب هرز و ۷۱۳۸ وب‌گاه (۷۹٪)، برچسب معتبر دارند. اطلاعاتی که از این مجموعه داده‌ای در اختیار شرکت‌کنندگان قرار گرفته است، گرافی شامل وب‌گاه‌های درون این مجموعه و پیوند بین آن‌ها، فایلی شامل اطلاعات مربوط به وزن TF-IDF کلمات درون وب‌گاه‌ها و

[۱] http://webspam.lip6.fr/wiki/pmwiki.php?n=Main.PhaseIITrainningCorpora/

همچنین برچسب وب‌گاه‌ها می‌باشد. گراف مربوط به این وب‌گاه‌ها به صورت جهت‌دار و وزن‌دار می‌باشد که وزن هر یال، تعداد ارجاعات بین هر زوج صفحه متمایز از دو وب‌گاه را نشان می‌دهد.

- مجموعه داده‌ای WEBSPAM-UK2007: این مجموعه داده‌ای در می ۲۰۰۷ توسط گروهی در دانشگاه میلان [۱۰۶] جمع‌آوری شده است. این مجموعه داده‌ای شامل ۱۱۴۵۲۹ وب‌گاه از دامنه uk. است که در مجموع دارای ۱۰۵۸۹۶۵۵ صفحه می‌باشند. از میان ۱۱۴۵۲۹ وب‌گاه موجود در این مجموعه داده‌ای، ۶۴۷۹ وب‌گاه برچسب‌گذاری شده‌اند که از میان آن‌ها ۵۷۰۹ وب‌گاه معتبر و ۳۴۴ وب‌گاه هرز می‌باشد. ۴۲۶ وب‌گاه باقی‌مانده نیز برچسب خط مرزی[۱] خورده‌اند که نشان‌گر صفحاتی است که نمی‌توان در مورد هرز بودن یا نبودن آن‌ها قضاوت کرد. آن چه که از این مجموعه داده‌ای در اختیار پژوهشگران قرار گرفته است، مجموعه‌ای از ویژگی‌های محتوایی و پیوندی و همچنین گراف بین وب‌گاه‌های این مجموعه داده‌ای می‌باشد. مجموعه ویژگی‌های محتوایی این مجموعه داده‌ای، شامل ۹۶ ویژگی می‌باشد که برای تعداد زیادی از وب‌گاه‌ها از قبل محاسبه شده‌اند. در میان این ویژگی‌های محتوایی تنها تعداد محدودی از ویژگی‌ها جزء ویژگی‌هایی هستند که در این پژوهش مورد بررسی قرار گرفته‌اند. بنابراین به دلیل عدم امکان محاسبه سایر ویژگی‌های محتوایی، امکان استفاده از این مجموعه داده‌ای برای ارزیابی روش‌های محتوایی ارائه شده در این پژوهش وجود ندارد. گراف این مجموعه داده‌ای نیز به صورت جهت‌دار و وزن‌دار بوده و وزن هر یال، تعداد ارجاعات صفحه یا صفحاتی از وب‌گاه مبدا به صفحه یا صفحاتی از وب‌گاه مقصد را مشخص می‌نماید.

۲.۴ معیارهای ارزیابی

برای ارزیابی کیفیت روش‌های ارائه شده در این پژوهش، به ازای هر یک از روش‌های شناسایی هرز وب، از معیارهای مشخصی استفاده شده است که در ادامه به معرفی هر یک از آن‌ها می‌پردازیم.

[۱] border line

۱.۲.۴ معیارهای ارزیابی وظیفه رده‌بندی هرز وب با استفاده از روش‌های مبتنی بر محتوا

به منظور ارزیابی روش‌هایی که برای رده‌بندی وب‌گاه‌های فارسی معرفی شده‌اند، از دو معیار دقت و فراخوانی برای هر دسته از صفحات هرز و معتبر استفاده شده است. برای محاسبه این دو معیار، ابتدا باید چهار پارامتر مثبت‌ـ‌صحیح[۱] (TP)، مثبت‌ـ‌کاذب[۲] (FP)، منفی‌ـ‌صحیح[۳] (TN) و منفی‌ـ‌کاذب[۴] (FN) را محاسبه کنیم. با توجه به این‌که هدف اصلی در این پژوهش شناسایی وب‌گاه‌های هرز می‌باشد، رخداد مثبت، هرز بودن یک وب‌گاه در نظر گرفته شده است. با توجه به این تعریف، یک رخداد مثبت‌ـ‌صحیح به شرایطی گفته می‌شود که برچسب حقیقی یک وب‌گاه و برچسبی که رده‌بند به آن داده است هر دو هرز باشند. به عبارت دیگر، یک صفحه هرز، توسط رده‌بند به درستی شناسایی شده باشد. همچنین رخداد مثبت‌ـ‌کاذب به شرایطی اطلاق می‌شود که یک وب‌گاه معتبر، به اشتباه توسط رده‌بند به عنوان یک وب‌گاه هرز برچسب‌گذاری شود. در صورتی که رده‌بند یک وب‌گاه معتبر را به درستی برچسب‌گذاری کند، حالت منفی‌ـ‌صحیح رخ داده است. حالت منفی‌ـ‌کاذب نیز به شرایطی گفته می‌شود که رده‌بند به اشتباه یک وب‌گاه هرز را معتبر تشخیص دهد. با توجه به این تعاریف، معیارهای دقت و فراخوانی برای وب‌گاه‌های هرز و معتبر به صورت زیر تعریف می‌شود:

$$Spam\_Precision = \frac{TP}{TP + FP} \quad (۱.۴)$$

$$Spam\_Recall = \frac{TP}{TP + FN} \quad (۲.۴)$$

$$NonSpam\_Precision = \frac{TN}{TN + FN} \quad (۳.۴)$$

[۱] true positive
[۲] false positive
[۳] true negative
[۴] false negative

$$Nonspam\_Recall = \frac{TN}{TN + FP} \quad (۴.۴)$$

لازم به ذکر است که، برخلاف بسیاری از روش‌های پیشین که این نکته را در نظر نگرفته‌اند، محاسبه این معیارها برای هر دو کلاس هرز و معتبر به صورت مجزا، امری ضروری می‌باشد. دلیل این امر، نامتوازن بودن تعداد وب‌گاه‌های هرز و معتبر است. به عبارت دیگر، نسبت تعداد وب‌گاه‌های معتبر به وب‌گاه‌های هرز در وب خیلی زیاد است. وجود این اختلاف باعث می‌شود که نتوانیم تحلیل درستی از میزان کارایی روش‌ها در رده‌بندی وب‌گاه‌ها به دو کلاس هرز و معتبر داشته باشیم. ضمن این‌که هدف اصلی ما شناسایی وب‌گاه‌های هرز می‌باشد که برای بررسی آن لازم است معیارهای دقت و فراخوانی برای این کلاس به طور مجزا محاسبه شود. همچنین، برای محاسبه تخمینی از کیفیت کلی رده‌بند، از معیار $F1$، که ترکیبی از دو معیار دقت و فراخوانی می‌باشد، استفاده شده است. مقدار این معیار، طبق رابطه ۵.۴ محاسبه می‌شود.

$$F1\_Score = \frac{2\ (Precision * Recall)}{(Precision + Recall)} \quad (۵.۴)$$

۲.۲.۴ معیارهای ارزیابی شناسایی هرز وب با استفاده از روش‌های مبتنی بر پیوند و روش ترکیبی

الگوریتم‌هایی که در بخش‌های ۲.۳ و ۳.۳ ارائه شده‌اند، در نهایت به هر صفحه یک امتیاز ارزشمندی یا عدم ارزشمندی اختصاص می‌دهند. سپس، صفحات بر اساس امتیازی که می‌گیرند رتبه‌بندی می‌شوند. در این پژوهش، برای ارزیابی این روش‌ها از دو نوع معیار ضریب هرز و ضریب اعتماد استفاده شده است. معیار ضریب هرز، که در مقاله [۷۳] معرفی شده است، برای ارزیابی روش‌های انتشار اعتماد استفاده می‌شود. در این الگوریتم‌ها، در صورتی که صفحات را به ترتیب امتیاز اعتمادی که بدست آورده‌اند به طور نزولی مرتب کنیم، هدف این است که صفحات معتبر در رتبه‌های بالاتر این فهرست قرار بگیرند. هر چقدر تعداد صفحات هرز در اوایل این فهرست کمتر باشد، میزان ضریب هرز نیز کمتر و توانایی الگوریتم در شناسایی هرز وب بیشتر است. با فرض این‌که صفحات به ترتیب s_1، s_2، s_3، ...، s_n رتبه‌بندی شده‌اند، ضریب هرز

به صورت زیر محاسبه می‌شود:

$$SpamFactor = \frac{\sum_{i=1}^{n} \omega(s_i) \times \frac{1}{i}}{\sum_{i=1}^{n} \frac{1}{i}}, \quad (۶.۴)$$

به طوری‌که در صورتی که صفحه s_i هرز باشد، مقدار $\omega(s_i)$ برابر با یک و در صورتی که معتبر باشد این مقدار برابر با صفر است.

همچنین برای سنجش میزان توانایی روش‌های انتشار هرز، معیاری به نام ضریب اعتماد را معرفی می‌نماییم که مشخص می‌کند چه میزان از صفحات هرز با چه دقتی توسط این الگوریتم‌ها شناسایی می‌شوند. برای محاسبه این معیار نیز ابتدا صفحات بر حسب امتیاز هرز بودنشان به طور نزولی مرتب می‌شوند. با فرض این‌که ترتیب صفحات به صورت p_1، p_2، p_3، ...، p_n باشد، ضریب اعتماد به صورت زیر محاسبه می‌شود:

$$ConfidenceFactor = \frac{\sum_{i=1}^{n} \omega(p_i) \times \frac{1}{i}}{\sum_{i=1}^{l} \frac{1}{i}}, \quad (۷.۴)$$

که در آن l تعداد کل صفحات هرز می‌باشد. در این عبارت، مخرج کسر حالت بهینه رتبه‌بندی صفحات می‌باشد که در آن تمام صفحات هرز در بالاترین رتبه ممکن قرار می‌گیرند.

۳.۴ نتایج آزمایش‌ها

در این بخش، نتایج آزمایش‌های صورت گرفته بر روی روش‌های معرفی شده به منظور شناسایی هرز وب را ارائه داده و با استفاده از معیارهای ارزیابی که در بخش ۲.۴ توضیح داده شد، نتایج را بررسی، تحلیل و

با یکدیگر مقایسه می‌نماییم.

۱.۳.۴ ارزیابی روش‌های مبتنی بر محتوا در شناسایی هرز وب فارسی

با توجه به این‌که هدف از این بخش، تحلیل روش‌های مبتنی بر محتوا بر روی وب‌گاه‌های فارسی می‌باشد، از مجموعه داده‌ای PersianWebSpam-2013 که به منظور انجام این پژوهش ساخته شده، استفاده شده است. صفحات موجود در این پیکره با استفاده از ابزار Lemur [۱۰۷] نمایه‌سازی شده و تمام ویژگی‌های محتوایی بیان شده در بخش ۱.۳ از آن‌ها استخراج شده است. پس از آن، با استفاده از ویژگی‌های محتوایی استخراج شده و همچنین ابزار Weka [۱۰۸]، وب‌گاه‌های موجود در این مجموعه داده‌ای رده‌بندی شده‌اند. همچنین برای رده‌بندی وب‌گاه‌ها در PSD-SYS از LIBSVM [۱۰۹]، که یک کتابخانه به زبان C++ است، استفاده کرده‌ایم. با توجه به این‌که روش کار ما در این بخش به صورت باسرپرست می‌باشد، برای افزایش دقت و اعتبار ارزیابی نتایج، از روش اعتبار سنجی متقاطع kـ بخشی[۱] استفاده شده است. بنابراین با در نظر گرفتن اندازه مجموعه داده‌ای، برای ارزیابی تمام روش‌های مبتنی بر محتوا از اعتبار سنجی ۵ـ بخشی استفاده کرده‌ایم.

ارزیابی مجموعه ویژگی‌های پایه

به عنوان یک روش پایه، ابتدا ویژگی‌های معرفی شده در مقاله [۸] را که در بخش ۲.۱.۳ به طور کامل توضیح دادیم استخراج کرده و با استفاده از انواع الگوریتم‌های یادگیری ماشین، وب‌گاه‌های فارسی موجود در پیکره را رده‌بندی کردیم. لازم به ذکر است که دو ویژگی مربوط به کلمات مشهور موجود در متن را، به ازای تعداد مختلف (۱۰۰، ۲۰۰، ۳۰۰، ۵۰۰ و ۱۰۰۰) کلمات مشهور محاسبه نمودیم. همچنین برای محاسبه مقدار دو ویژگی شباهت nـ گرام‌ها، مقادیر مختلف ۱، ۲، ۳، ۴ و ۵ را برای احتمال مستقل و مقادیر ۲، ۳، ۴ و ۵ را برای احتمال وابسته در نظر گرفتیم. نتایج رده‌بندهای مختلف با استفاده از مجموعه ویژگی‌های پایه در جدول ۱.۴ ارائه شده است.

[۱] k-fold cross validation

جدول ۱.۴: نتایج استفاده از الگوریتم‌های یادگیری ماشین برای رده‌بندی وب‌گاه‌های PersianWebSpam-2013 با استفاده از مجموعه ویژگی‌های پایه

NonSpam			Spam			
F1	Recall	Precision	F1	Recall	Precision	رده‌بند
۸۴	۸۴/۸۶	**۸۳/۲**	۴۱/۴۵	۴۰	۴۳/۳۳	**KNN**
۸۱/۲۲	۸۰/۲۹	۸۲/۵۳	۳۸/۳۹	**۴۰/۳۳**	۳۷/۹۹	**Naive Bayes**
۸۷/۸	۹۴/۷۶	۸۱/۸۷	۳۵/۷۷	۲۶/۳۳	۶۱/۳۷	**Logistic Regression**
۸۶/۳۳	۹۰/۲۹	۸۲/۸۱	۴۰/۶۵	۳۴/۳۳	۵۲/۱	**C4.5**
۸۷/۹۱	**۹۸/۶۷**	۷۹/۲۸	۱۶/۳۱	۹/۶۷	**۷۰/۲۵**	**SVM**
۸۶/۷۷	۹۰/۹۵	۸۲/۹۹	**۴۱/۵۱**	۳۴/۶۷	۵۲/۳۶	**Random Forest**

با توجه به جدول ۱.۴ مشاهده می‌نماییم که روش جنگل تصادفی بیشترین امتیاز $F1$ را برای کلاس هرز دارد. با توجه به اهمیت زیاد رده‌بندی وب‌گاه‌های هرز در این وظیفه، در ادامه برای رده‌بندی وب‌گاه‌ها از الگوریتم جنگل تصادفی استفاده می‌شود. جنگل تصادفی یک رده‌بند تجمیعی است که تعداد زیادی درخت تصمیم ایجاد کرده و برای برچسب‌گذاری یک نمونه آزمون، از رای اکثریت استفاده می‌کند. با استفاده از این روش از فرابرازش داده‌ها جلوگیری شده و برچسب‌گذاری نمونه‌ها، تحت تاثیر داده‌های برون‌هشته قرار نمی‌گیرد. بنابراین استفاده از این روش برای رده‌بندی وب‌گاه‌های هرز که دارای مقادیر برون‌هشته زیادی برای ویژگی‌های مختلف هستند مناسب می‌باشد.

ارزیابی مجموعه ویژگی‌های مکمل

با توجه به این‌که صفحات وب از بخش‌های مختلفی تشکیل شده‌اند، هر کدام دارای خصوصیات زیادی هستند که می‌توان با استفاده از انواع ویژگی‌ها آن‌ها را شناسایی کرد. بنابراین در این بخش، ویژگی‌های مکملی که به مرور زمان در پژوهش‌های مختلف [۲۱، ۲۹، ۱۰۰] معرفی شده است را از وب‌گاه‌های فارسی استخراج کرده و با استفاده از الگوریتم جنگل تصادفی وب‌گاه‌ها را رده‌بندی کردیم. نتایج این رده‌بندی در جدول ۲.۴ ارائه شده است.

جدول ۲.۴: نتایج استفاده از مجموعه ویژگی‌های پایه، مکمل و جدید برای رده‌بندی وب‌گاه‌های موجود در مجموعه داده‌ای PersianWebSpam-2013

ویژگی‌ها	Spam			NonSpam		
	Precision	Recall	F1	Precision	Recall	F1
مجموعه ویژگی پایه	۵۲/۳۶	۳۴/۶۷	۴۱/۵۱	۸۲/۹۹	۹۰/۹۵	۸۶/۷۷
مجموعه ویژگی پایه و مکمل	۶۸/۹۹	۴۸/۶۷	۵۶/۹۳	۸۶/۵۱	۹۳/۸۱	۹۰
مجموعه ویژگی پایه و جدید	۶۶/۵۶	۴۹	۵۶/۲۴	۸۶/۴۳	۹۲/۷۶	۸۹/۴۷
مجموعه ویژگی پایه، مکمل و جدید	**۶۹/۸**	**۴۹**	**۵۷/۴۹**	**۸۶/۵۵**	**۹۳/۸۱**	**۹۰/۰۳**

با توجه به نتایج موجود در جدول ۲.۴، استفاده از ویژگی‌های مکمل، نتایج رده‌بندی محتوایی را از نظر معیار $F1$ برای کلاس هرز ۳۷/۱۵٪ و برای کلاس معتبر ۳/۷۲٪ بهبود داده است.

ارزیابی مجموعه ویژگی‌های جدید

در این بخش، برای ارزیابی ویژگی‌های محتوایی جدیدی که در این پژوهش معرفی شده‌اند، آن‌ها را به مجموعه ویژگی‌های پایه اضافه کرده و نتایج رده‌بندی را بررسی می‌نماییم. نتایج این رده‌بندی در جدول ۲.۴ ارائه شده است. با توجه به نتایج مشاهده می‌کنیم که با استفاده از این ویژگی‌های جدید محتوایی می‌توان نتایج رده‌بندی را به نسبت ویژگی‌های پایه ۳۵/۴۹٪ از نظر معیار $F1$ برای کلاس هرز و ۳/۱۱٪ برای کلاس معتبر بهبود داد. همان‌طور که واضح است این میزان بهبود بسیار نزدیک به میزان بهبودی است که با استفاده از ویژگی‌های مکمل در نتایج رده‌بندی ایجاد شده است. این امر نشان می‌دهد که میزان کارایی ویژگی‌های محتوایی جدیدی که در این پژوهش معرفی شده‌اند برابر با کارایی مجموعه ویژگی‌های مکملی است که تاکنون در پژوهش‌های مختلف ارائه شده است. این میزان بهبود در حالی است که تعداد ویژگی‌های محتوایی جدید تقریبا نصف تعداد ویژگی‌های مکمل است. همچنین هزینه محاسباتی این ویژگی‌های جدید نیز بسیار کمتر از هزینه محاسباتی بسیاری از ویژگی‌های پایه و مکمل می‌باشد.

پس از آن، ویژگی‌های محتوایی جدید را با دو مجموعه ویژگی پایه و مکمل ترکیب کرده و نتایج رده‌بندی را بررسی می‌نماییم. با توجه به جدول ۲.۴ مشاهده می‌کنیم که ترکیب ویژگی‌های جدید با دو

مجموعه ویژگی پایه و مکمل، نتایج رده‌بندی را از نظر معیار $Spam\_F1$ حدود ۰/۹۸٪ بهبود داده است. همچنین این رده‌بند نهایی در مقایسه با رده‌بند پایه، ۳۸/۵٪ بهبود در معیار $Spam\_F1$ و ۳/۷۶٪ بهبود در معیار $NonSpam\_F1$ ایجاد کرده است.

اگرچه ترکیب ویژگی‌های جدید با دو مجموعه ویژگی پایه و مکمل، نتایج رده‌بندی وب‌گاه‌های فارسی را بهبود داده است اما استخراج تمام این ویژگی‌ها از مجموعه وب‌گاه‌ها، نیاز به صرف هزینه‌های محاسباتی و زمانی زیادی دارد. همچنین، ترکیب برخی از این ویژگی‌ها با یکدیگر باعث کاهش دقت رده‌بندی می‌شود. بنابراین برای رفع این مشکلات در ادامه از روش انتخاب ویژگی χ^2 -test استفاده شده است.

نتایج روش انتخاب ویژگی و رده‌بندی نهایی

در این مرحله، برای بهبود نتایج رده‌بندی و کاهش هزینه ابتدا، از روش χ^2 -test برای امتیازدهی به ویژگی‌ها استفاده شده است. پس از آن برای انتخاب زیرمجموعه‌ای از ویژگی‌های استخراج شده به عنوان ویژگی‌های بهینه، ابتدا تمام ویژگی‌ها را بر اساس امتیازی که از χ^2 -test بدست آورده‌اند، به صورت نزولی مرتب کرده و از روش حذف پس‌رو برای حذف ویژگی‌های غیربهینه استفاده کردیم. با استفاده از این روش، ابتدا رده‌بندی را با تمام ویژگی‌های استخراج شده از صفحات انجام داده و در هر مرحله به ترتیب، ویژگی با کمترین امتیاز χ^2 -test را از مجموعه ویژگی‌ها حذف کردیم. در صورتی که حذف ویژگی جدید باعث کاهش کیفیت نتایج رده‌بندی شده آن ویژگی را مجددا به مجموعه ویژگی‌های باقی‌مانده اضافه کردیم، در غیر این صورت پس از حذف آن ویژگی، ویژگی بعدی را بررسی نمودیم. بدین ترتیب در نهایت تعدادی ویژگی به عنوان ویژگی‌های بهینه انتخاب شدند. این مجموعه ویژگی‌ها به همرا امتیاز χ^2 -test در جدول ۳.۴ ارائه شده‌اند.

همان‌طور که در جدول ۳.۴ مشاهده می‌نماییم، از ۷ ویژگی معرفی شده در این پژوهش، ۵ ویژگی آن جزء مجموعه ویژگی‌های بهینه انتخاب شده است.

در بین این ویژگی‌های بهینه، برای مشخص کردن تاثیر هر ویژگی در رده‌بندی، لازم است وزنی که رده‌بند به هریک از آن‌ها می‌دهد را محاسبه نماییم. همان‌طور که در جدول ۱.۴ مشاهده کردیم، الگوریتم

جدول ۳.۴: ویژگی‌های بهینه در شناسایی وب‌گاه‌های هرز فارسی

ویژگی	امتیاز χ^2	رتبه
طول URL	۱۹۴/۸۸۴	۱
تعداد پیوندهای خروجی	۹۳/۲۸۹	۲
احتمال شباهت ۴ـ گرام‌های شرطی	۸۱/۶۲۳	۳
شباهت کسینوسی بین بدنه اصلی و متن پیوند با وزن‌دهی TF	۷۵/۸۰۲	۴
احتمال شباهت ۳ـ گرام‌های شرطی	۶۱/۴۹۷	۵
درصدی از صفحه که شامل ایست‌واژه‌ها است	۵۹/۱۸۷	۶
شباهت کسینوسی بین عنوان و متن پیوند با وزن‌دهی TF	۵۸/۷۷۲	۷
درصد فشرده‌سازی	۵۶/۷۹۷	۸
شباهت کسینوسی بین متن پیوند و برچسب کلیدواژه‌ها و توضیحات با وزن‌دهی TF	۵۵/۵۸۵	۹
تعداد i-frameها	۵۴/۳۹	۱۰
اندازه متن پیوند	۵۲/۴۶۷	۱۱
درصدی از صفحه که شامل ۱۰۰ کلمه مشهور پیکره است	۳۹/۷۲۹	۱۲
احتمال شباهت ۵ـ گرام‌های شرطی	۳۹/۴۰۸	۱۳
تعداد منابع چندرسانه‌ای	۳۹/۲۱۴	۱۴
تعداد عکس‌ها	۳۹/۲۰۹	۱۵
تعداد ابربرچسب‌های جاوا اسکریپت	۳۷/۸۴۲	۱۶
تعداد کلمات عنوان	۳۴/۴۳۶	۱۷
درصد متن پیوند درون صفحه	۳۴/۳۴	۱۸
تعداد کلمات درون ابربرچسب کلیدواژه‌ها و توضیحات	۳۳/۷۷۲	۱۹
شباهت کسینوسی بین بدنه اصلی و برچسب کلیدواژه‌ها و توضیحات با وزن‌دهی TF	۳۰/۶۲۴	۲۰
متوسط طول کلمات	۲۶/۳۶	۲۱
تعداد کلمات بدنه اصلی	۱۹/۵۴۶	۲۲
تعداد کلمات صفحه	۱۹/۴۸۱	۲۳
احتمال شباهت ۵ـ گرام‌های مستقل	۱۶/۲۲۶	۲۴

جنگل تصادفی بیشترین میزان $Spam\_F1$ را در رده‌بندی وب‌گاه‌های فارسی با استفاده از ویژگی‌های معرفی شده دارد. همان‌طور که می‌دانیم در الگوریتم جنگل تصادفی تمام حالت‌های انتخاب ویژگی‌ها بر اساس اهمیتشان بررسی می‌شود و برچسب نمونه‌ها بر اساس رای اکثریت درخت‌ها مشخص می‌شود. در واقع این یکی از مزیت‌های الگوریتم جنگل تصادفی برای وظیفه شناسایی هرز وب می‌باشد، که تمام حالت‌های مختلف انتخاب زیر مجموعه‌ای از ویژگی‌ها و ترتیب استفاده از آن‌ها را در نظر می‌گیرد و به هر ویژگی به تنهایی یک وزن مشخصی نمی‌دهد. همان‌طور که Ntoulas و همکاران نیز در مقاله [۸] بیان کرده‌اند، خصوصیات صفحات وب و ویژگی‌های محتوایی تعریف شده طوری می‌باشد که هر ویژگی به تنهایی تاثیر زیادی در شناسایی برچسب صفحات ندارد و برای محاسبه تاثیر آن‌ها لازم است تمام ویژگی‌ها در کنار هم در نظر گرفته شوند. با این حال اگر بخواهیم میزان تاثیر هر ویژگی را به طور مجزا در فرآیند رده‌بندی مشخص نماییم، می‌توان رده‌بندی را با استفاده از هر یک از ویژگی‌ها به طور مجزا انجام داده و با توجه به مقدار $Spam\_F1$، میزان تاثیر هر ویژگی را در شناسایی وب‌گاه‌های هرز فارسی بررسی نماییم. با استفاده از این روش، ۱۵ ویژگی برتر به ترتیب بر اساس مقدار $Spam\_F1$ در جدول ۴.۴ ارائه شده‌اند.

با توجه به جدول ۴.۴، موثرترین ویژگی محتوایی در رده‌بندی وب‌گاه‌های فارسی، طول URL وب‌گاه‌ها می‌باشد. این امر بیانگر این نکته است که هرزنویسان برای ایجاد وب‌گاه‌های هرز فارسی از URLهایی با تعداد زیادی کلیدواژه استفاده می‌کنند. دومین ویژگی موثر در شناسایی وب‌گاه‌های هرز، احتمال شرطی ۳ـ گرام‌ها می‌باشد. این ویژگی نشان می‌دهد که تعداد زیادی از وب‌گاه‌های هرز با استفاده از تکرار عبارت‌های ۳ـ گرامی تلاش می‌کنند رتبه‌شان را افزایش دهند. برای مثال عبارت «مدل لباس جدید» در بسیاری از وب‌گاه‌های هرز به تعداد زیاد تکرار شده است. ویژگی سوم، تعداد پیوندهای خروجی می‌باشد. این امر نشان‌دهنده این مهم می‌باشد که وب‌گاه‌های هرز فارسی برای افزایش رتبه وب‌گاه‌های خود، تعداد زیادی از پیوندهای سایر وب‌گاه‌ها را درون صفحات خود قرار می‌دهند. ویژگی بعد، تعداد کلمات درون متن پیوند است که به طور متوسط در وب‌گاه‌های هرز بیشتر از وب‌گاه‌های معتبر می‌باشد. این وب‌گاه‌ها برای بالا بردن رتبه خود در میان نتایج پرس‌وجوها، تعداد زیادی ارجاع به صفحات مختلف می‌دهند که هر کدام از آن‌ها شامل توضیحاتی پیرامون صفحه ارجاع داده شده و کلیدواژه‌های مربوط به آن‌ها می‌باشند.

جدول ۴.۴: بررسی کارایی ویژگی‌های بهینه درشناسایی وب‌گاه‌های هرز

ویژگی	Spam-F1	رتبه
طول URL	۳۰/۸	۱
احتمال شباهت ۳ـ گرام‌های شرطی	۲۹/۱	۲
تعداد پیوندهای خروجی	۲۷/۷	۳
اندازه متن پیوند	۲۶/۹	۴
درصدی از صفحه که شامل ۱۰۰ کلمه مشهور پیکره است	۲۶/۹	۵
احتمال شباهت ۵ـ گرام‌های شرطی	۲۶/۵	۶
تعداد کلمات صفحه	۲۶/۲	۷
احتمال شباهت ۴ـ گرام‌های شرطی	۲۵/۹	۸
شباهت کسینوسی بین عنوان و متن پیوند با وزن‌دهی TF	۲۵/۸	۹
شباهت کسینوسی بین بدنه اصلی و متن پیوند با وزن‌دهی TF	۲۵/۸	۱۰
درصد فشرده‌سازی	۲۵/۷	۱۱
شباهت کسینوسی بین بدنه اصلی و برچسب کلیدواژه‌ها و توضیحات با وزن‌دهی TF	۲۴/۷	۱۲
درصدی از صفحه که شامل ایست‌واژه‌ها است	۲۴/۶	۱۳
تعداد کلمات بدنه اصلی	۲۴/۳	۱۴
تعداد منابع چندرسانه‌ای	۲۳/۹	۱۵

جدول ۵.۴: نتایج استفاده از χ^2 -test در رده‌بندی وب‌گاه‌های PersianWebSpam-2013

ویژگی‌ها	Spam			NonSpam		
	Precision	Recall	F1	Precision	Recall	F1
مجموعه ویژگی پایه، مکمل و جدید	۶۹/۸	۴۹	۵۷/۴۹	۸۶/۵۵۲	۹۳/۸۱	۹۰/۰۲۹
مجموعه ویژگی‌های بهینه	۷۲/۴۸	۵۲/۶۷	۶۰/۹	۸۷/۴۵	۹۴/۱۹	۹۰/۶۹

پس از انتخاب مجموعه ویژگی‌های بهینه و حذف سایر ویژگی‌ها، نتایج رده‌بندی با استفاده از مجموعه ویژگی‌های بهینه بدست آمده است. با توجه به جدول ۵.۴ می‌توان مشاهده کرد که حذف ویژگی‌های غیربهینه از بردار ویژگی، نتایج رده‌بندی را به اندازه ۵/۹٪ از نظر معیار $Spam\_F1$ و ۰/۷٪ از لحاظ معیار $NonSpam\_F1$ بهبود داده است. علاوه بر این امر، در این روش با توجه به کاهش تعداد ویژگی‌ها، هزینه زمانی و محاسباتی کل فرآیند رده‌بندی نیز کاهش می‌یابد.

ارزیابی رده‌بندی وب‌گاه‌های فارسی با استفاده از PSD-SYS

در این سامانه، پس از حذف ایست‌واژه‌ها و کلمات مربوط به ابربرچسب‌ها و نمایه‌سازی صفحات، کلمات رایج موجود در وب‌گاه‌های هرز مشخص شده و برای رده‌بندی وب‌گاه‌ها از مدل BOSW استفاده شده است. تعداد ویژگی‌هایی که در این سامانه استخراج و استفاده می‌شود حدود ۳۵۰۰۰ ویژگی است که تعداد آن در مقایسه با تعداد وب‌گاه‌ها بسیار کمتر می‌باشد. در این بردار ویژگی، به ازای هر ویژگی مقدار صفر یا یک وجود دارد. همچنین پس از بررسی تعداد زیادی از الگوریتم‌های یادگیری ماشین، در نهایت برای رده‌بندی وب‌گاه‌ها از الگوریتم SVM استفاده شده است. بدین منظور از ابزار LIBSVM [۱۰۹] با پارامترهای $c = 10$ و $t = 2$ استفاده شده است. جدول ۶.۴ نتایج رده‌بندی وب‌گاه‌های موجود در مجموعه داده‌ای WebSpamPersian-2013 با استفاده از مدل BOSW را با نتایج رده‌بندی آن‌ها با استفاده از مدل ساده کیف کلمات که در پژوهش‌های پیشین [۹۰، ۹۶، ۱۱۰] معرفی شده است، مقایسه می‌کند.

با توجه به جدول ۶.۴، استفاده از روش BOSW برای رده‌بندی وب‌گاه‌های فارسی نتایج رده‌بندی را به

جدول ۶.۴: مقایسه رده‌بندی وب‌گاه‌های مجموعه داده‌ای WebSpamPersian-2013 با استفاده از مدل BOSW و مدل کیف کلمات

روش	Spam			NonSpam		
	Precision	Recall	F1	Precision	Recall	F1
کیف کلمات	۸۶/۵۳	۸۶/۳۳	۸۶/۳۵	۹۶/۱	۹۶/۰۹	۹۶/۰۹
BOSW	۹۲/۶۱	۸۶	۸۹/۱۲	۹۶/۰۸	۹۸	۹۷/۰۳

نسبت روش کیف کلمات ۳/۲۱٪ از نظر معیار $Spam\_F1$ و ۰/۹۸٪ از نظر معیار $NonSpam\_F1$ بهبود داده است. با توجه به این نتایج مشاهده می‌کنیم که استفاده از روش BOSW برای شناسایی وب‌گاه‌های فارسی بهتر از روش کیف کلمات می‌باشد. دلیل این امر این است که وب‌گاه‌های معتبر دارای موضوعات متنوعی هستند که در مجموع کلماتی که در این نوع از وب‌گاه‌ها هستند دارای پراکندگی زیاد و مربوط به موضوعات مختلف هستند. همان‌طور که توضیح داده شد، در روش کیف کلمات، تمام این مجموعه کلمات به همراه کلمات مربوط به وب‌گاه‌های هرز، به عنوان بردار ویژگی رده‌بندی انتخاب می‌شوند. انتخاب تمام این کلمات به عنوان ویژگی‌های رده‌بندی، علاوه بر افزایش هزینه رده‌بندی، باعث ایجاد خطاهایی در رده‌بندی می‌شود. از طرف دیگر، وب‌گاه‌های هرز معمولا دارای کلماتی پیرامون موضوعات محدود مانند کلمات تبلیغاتی مربوط به کالاهای تقلبی خاص یا کلمات مربوط به فروش محصولات و کلماتی از این قبیل می‌باشند که معمولا ترکیب خاصی از این کلمات در وب‌گاه‌های هرز مشاهده می‌شود. برای مثال در وب‌گاه‌های فارسی موجود در مجموعه داده‌ای WebSpamPersian-2013، کلمات «خرید»، «محصول»، «اس‌ام‌اس»، «دانلود»، «تخفیف»، «ویژه»، «رایگان» و کلماتی از این قبیل بسیار رایج هستند. استفاده از این کلمات برای شناسایی وب‌گاه‌های هرز روش مناسبی می‌باشد. در صورتی که اگر کلمات وب‌گاه‌های معتبر نیز به این بردار ویژگی اضافه شوند، باعث می‌شود تعداد زیادی از وب‌گاه‌های معتبر به اشتباه به عنوان هرز شناسایی شوند. برای مثال در صورتی که یک وب‌گاه معتبر مورد سوءاستفاده وب‌گاه‌های هرز قرار بگیرد و متن‌های این وب‌گاه، درون تعدادی از وب‌گاه‌های هرز رونوشت شود، الگوی مربوط به کلمات درون این وب‌گاه معتبر به عنوان مشخصه‌ای از وب‌گاه‌های هرز شناسایی شده و به آن صفحه به اشتباه برچسب هرز داده می‌شود.

همان‌طور که می‌دانیم برای رده‌بندی نمونه‌ها توسط الگوریتم SVM از بردار ویژگی استفاده می‌شود. در سامانه PSD-SYS، این بردار شامل مجموعه‌ای از کلمات رایج در مجموعه صفحات هرز می‌باشد. این روش بر اساس این فرض ارائه شده است که صفحات هرز دارای مجموعه‌ای از کلمات هرز هستند که رخداد ترکیب خاصی از این کلمات در یک صفحه نشان‌گر هرز بودن آن صفحه است. احتمال وجود این ترکیب‌های خاص کلمات هرز، در صفحات معتبر بسیار کم می‌باشد. برای مثال کلمه‌ای مانند «کنکور» ممکن است هم در صفحات هرز هم معتبر وجود داشته‌باشد. وجود این کلمه در کنار کلماتی مانند «دست‌بند»، «خرید»، «لاغری»، «عکس»، «اس‌ام‌اس» و «گیاهی» که جزء کلمات هرز هستند، نشان‌گر یک صفحه هرز می‌باشد. در صورتی که اگر این کلمه به همراه کلماتی مانند «خواندن»، «برنامه‌ریزی»، «رشته»، «دانشگاه» و کلمات دیگری که جزء کلمات رایج هرز نمی‌باشند در صفحه ظاهر شود، این صفحه به عنوان یک صفحه معتبر شناسایی می‌شود. در رده‌بند SVM که در این پژوهش از آن برای رده‌بندی وب‌گاه‌ها استفاده شده است، برخی از این ترکیب‌های خاص، همان بردارهای پشتیبان[۱] هستند که برای جدا کردن فضای صفحات هرز و معتبر استفاده می‌شوند. در صورتی که با همان احتمال کم، تعدادی از صفحات معتبر دارای ترکیبی خاص از کلمات هرز باشند، به طوری‌که در رده‌بندی این صفحه با استفاده از بردار ویژگی مربوط به آن، رده‌بند نتواند فرقی بین این صفحه و صفحات هرز قائل شود، این صفحه به عنوان صفحه هرز برچسب گذاری می‌شود. برای مثال، ممکن است در بخش نظرات یک صفحه معتبر، هرزنویسان آدرس صفحات خود را به همراه توضیحاتی درباره صفحه خود که شامل کلمات تبلیغاتی زیادی است قرار داده باشند و تعداد این کلمات و ترکیب آن‌ها طوری باشد که احتمال هرز بودن آن صفحه بیشتر از اعتبار آن شود و در نتیجه صفحه به عنوان یک صفحه هرز شناسایی شود. اما همان‌طور که در نتایج جدول ۶.۴ مشاهده می‌نماییم، احتمال وجود چنین صفحاتی کم است. به عبارت دیگر $Spam\_F1$ برای این روش برابر با ۹۲/۶۱ می‌باشد، بدین معنا که تنها حدود ۷/۵ درصد از صفحات معتبر با استفاده از این روش به اشتباه به عنوان هرز برچسب خورده‌اند. همان‌طور که در نتایج جدول ۶.۴ مشخص است، این مقدار خطا حدود نصف خطای موجود در روش کیف کلمات (با خطای ۱۳/۵ درصد) است.

در ادامه، در این سامانه برای کاهش تعداد ویژگی‌ها و کاهش هزینه‌های زمانی و محاسباتی از

[۱] support vectors

جدول ۷.۴: نتایج استفاده از روش‌های مختلف انتخاب ویژگی در PSD-SYS

روش	Spam			NonSpam		
	Precision	Recall	F1	Precision	Recall	F1
انتخاب ۲ ـ گرام‌ها بدون محدودیت	۸۸/۸۳	۸۵/۶۷	۸۷/۱۳	۹۵/۹۵	۹۶/۸۶	۹۶/۴
انتخاب ۲ ـ گرام‌ها با تعداد تکرار ۵ و بیشتر	۹۳/۲۶۲	۸۶/۳۳۳	۸۹/۶۰۱	۹۶/۱۸۶	۹۸/۱۹	۹۷/۱۷۳
Mutual Information	۹۰/۹۶۷	۸۵/۳۳۳	۸۷/۹۹۳	۹۵/۸۸۶	۹۷/۵۲۴	۹۶/۶۹۳
χ^2 **-test**	۹۱/۵۱۹	۸۷/۶۶۷	۸۹/۴۶۷	۹۶/۵۲۶	۹۷/۶۱۹	۹۷/۰۶۳
انتخاب ۱ ـ گرام‌ها با تعداد تکرار ۴ و بیشتر	۹۳/۴۵۴	۸۸/۳۳۳	۹۰/۷۶۷	۹۶/۷۲۲	۹۸/۱۹	۹۷/۴۴۶
انتخاب ۱ ـ گرام‌ها بدون محدودیت	۹۲/۶۰۸	۸۶	۸۹/۱۲۴	۹۶/۰۸۵	۹۸	۹۷/۰۲۹
انتخاب ۱ ـ گرام‌ها با TF-IDF بزرگتر و یا مساوی ۲۰	۹۲/۶۷۵	۸۶/۶۶۷	۸۹/۵۲	۹۶/۲۶۲	۹۸	۹۷/۱۱۹

روش‌های مختلف انتخاب ویژگی استفاده شده است. نتایج اعمال روش‌های مختلف انتخاب بردار ویژگی در جدول ۷.۴ ارائه شده است. هر روش، به ازای مقادیر آستانه متفاوت آزمایش شده و نتیجه مربوط به مقدار بهینه آن در جدول گزارش شده است.

با توجه به این نتایج مشاهده می‌نماییم که استفاده از روش وزن‌دهی TF و انتخاب ۱ ـ گرام‌های رایج در وب‌گاه‌های هرز با تعداد تکرار بزرگتر و مساوی ۴ به عنوان بردار ویژگی، بهترین نتایج را از لحاظ معیار $F1$ برای هر دو کلاس هرز و معتبر دارد. با مقایسه نتایج این جدول با نتایج جدول ۶.۴ مشاهده می‌شود که روش انتخاب ویژگی و حذف ویژگی‌های غیربهینه، نتایج رده‌بندی را از لحاظ دو معیار $Spam\_F1$ و $NonSpam\_F1$ بهبود داده است.

در نهایت، مقایسه نتایج حاصل از رده‌بندی وب‌گاه‌ها با استفاده از PSD-SYS با نتایج جدول ۵.۴، نشان می‌دهد که استفاده از PSD-SYS برای رده‌بندی وب‌گاه‌ها در مقایسه با روش محتوایی بخش ۲.۱.۳، نتایج رده‌بندی را ۴۶/۰۵٪ از لحاظ معیار $Spam\_F1$ و ۷/۴۵٪ از لحاظ معیار $NonSpam\_F1$ بهبود داده است. با توجه به این نتایج مشاهده می‌کنیم که استفاده از ویژگی‌های محتوایی مربوط به ساختار وب‌گاه‌ها در رده‌بندی آن‌ها کارایی بالایی ندارد. دلیل این امر، وجود انواع مختلفی از وب‌گاه‌های معتبر با خصوصیات متفاوت است که موجب می‌شود مرز خصوصیات ساختاری بین وب‌گاه‌های معتبر و هرز

کم‌رنگ شود. همچنین امروزه بسیاری از هرزنویسان برای مقابله با انواع روش‌های شناسایی هرز وب تلاش می‌کنند صفحات هرزی را ایجاد کنند که از لحاظ بسیاری از خصوصیات ساختاری، مشابه وب‌گاه‌های معتبر هستند. آن‌ها با استفاده از این روش سعی دارند روش‌های محتوایی شناسایی هرز وب را فریب دهند. از طرف دیگر، برای بالا بردن رتبه وب‌گاه‌های خود در میان نتایج پرس‌وجوها، از تعداد زیادی کلیدواژه‌های رایج در پرس‌وجوها استفاده می‌کنند. همان‌طور که نتایج نشان می‌دهد، این نوع از وب‌گاه‌های هرز با استفاده از PSD-SYS قابل شناسایی می‌باشند.

۲.۳.۴ ارزیابی روش‌های مبتنی بر پیوند در شناسایی هرز وب

در این بخش، نتایج مربوط به رتبه‌بندی وب‌گاه‌ها با استفاده از هر یک از الگوریتم‌های مبتنی بر پیوند WorthyRank و JunkyRank ارائه داده می‌شود. برای ارزیابی این روش‌ها، از دو مجموعه داده‌ای استاندارد WebSpamChallengeII-CorpusI و WEBSPAM-UK2007 استفاده شده است. برای تحلیل میزان کارایی الگوریتم‌های معرفی شده، نتایج آن‌ها را با نتایج حاصل از تعدادی از روش‌های پیشین در این زمینه مقایسه می‌نماییم. بدین منظور، ابتدا به توضیح مختصر هر یک از روش‌های پایه پرداخته و سپس عملکرد آن‌ها را با عملکرد الگوریتم‌های معرفی شده در این پژوهش مقایسه می‌کنیم.

معرفی روش‌های پایه

در این بخش، ابتدا تعدادی از روش‌های پیشین که به عنوان روش پایه این الگوریتم‌ها در نظر گرفته شده‌اند را به اختصار توضیح می‌دهیم.

- الگوریتم TrustRank: این الگوریتم که برای اولین بار توسط Gyongyi و همکاران [۱۷] معرفی شده است، بر اساس این فرض است که صفحات معتبر در وب، با احتمال زیادی به صفحات معتبر ارجاع می‌دهند. بنابراین در این الگوریتم، امتیاز اعتماد صفحات از مجموعه‌ای از صفحات معتبر که به عنوان صفحات بذر انتخاب شده‌اند، در جهت یال‌های گراف به سایر صفحات گراف انتشار داده

می‌شود. همچنین در هر تکرار الگوریتم، صفحات بذر علاوه بر امتیازی که از پیوندهای ورودی از صفحات همسایه کسب می‌کنند، وزن ثابتی از امتیاز اولیه‌شان را نیز دریافت می‌کنند. بدین ترتیب در نهایت برای هر گره در گراف، یک امتیاز اعتماد محاسبه می‌شود. رابطه ۸.۴ نحوه انتشار و محاسبه این امتیاز را برای هر گره موجود در گراف وب نشان می‌دهد.

$$\vec{t} = \alpha * M^T * \vec{t} + (1-\alpha) * \vec{s} \qquad (۸.۴)$$

در این رابطه، $\vec{t}$ بردار امتیاز اعتماد، M ماتریس مجاورت به‌هنجار شده گراف وب، و α عامل میرایی می‌باشد. بردار s نیر بردار به‌هنجار شده مقدار اولیه امتیاز اعتماد صفحات بذر است که در آن، در صورتی که p عضوی از مجموعه صفحات بذر معتبر باشد $s(p) = \frac{1}{|S^+|}$، و در غیر این صورت برابر با صفر است.

- الگوریتم Anti-TrustRank: فرض دیگری که بعدها برای رابطه بین صفحات وب در نظر گرفته شد، این است که صفحاتی که به یک صفحه هرز ارجاع داده‌اند با احتمال زیادی هرز هستند. با توجه به این قانون، الگوریتم Anti-TrustRank [۶۰] مطرح شد که در آن، امتیاز هرز بودن مجموعه‌ای از صفحات بذر هرز، در خلاف جهت یال‌های گراف به سایر صفحات گراف انتشار داده می‌شود. در این الگوریتم نیز مانند الگوریتم TrustRank به هر صفحه بذر، علاوه بر امتیازی که از سایر صفحات دریافت می‌کند، درصدی از امتیاز ثابت هرز بودن اختصاص داده می‌شود. نحوه انتشار امتیاز هرز در گراف وب با استفاده از این الگوریتم در رابطه ۹.۴ مشخص شده است.

$$\vec{d} = \alpha * M * \vec{d} + (1-\alpha) * \vec{s'} \qquad (۹.۴)$$

در این رابطه، $\vec{d}$ بردار امتیاز هرز بودن، M ماتریس مجاورت به‌هنجار شده گراف، و α عامل میرایی می‌باشد. $\vec{s'}$ نیز بردار به‌هنجار شده مقدار اولیه امتیاز هرز بودن صفحات بذر است که در آن، در صورتی که p عضوی از مجموعه صفحات بذر هرز باشد $s'(p) = \frac{1}{|S^-|}$، و در غیر این صورت برابر

با صفر است.

- الگوریتم TDR: این الگوریتم که در سال ۲۰۱۱ توسط Zhang و همکاران [۷۲] معرفی شده است، از انتشار هم‌زمان دو امتیاز اعتماد و عدم اعتماد در گراف وب استفاده می‌کند. فرض آن‌ها بر این است که هر صفحه می‌تواند هم‌زمان دارای دو امتیاز هرز و اعتبار به صورت مجزا از هم باشد. همچنین در این الگوریتم برای انتشار امتیاز از گره i به گره j، احتمال اعتبار یا هرز بودن گره مقصد در نظر گرفته می‌شود. نویسندگان این مقاله عقیده دارند که میزان اعتبار (عدم اعتبار) یک صفحه، در مقدار وزن اعتباری (عدم اعتبار) که از سایر صفحات وب دریافت می‌کند تاثیر دارد. در این الگوریتم در هر بار تکرار، امتیاز اعتبار یا هرز بودن صفحات به ترتیب با استفاده از روابط ۱۰.۴ و ۱۱.۴ محاسبه می‌شود.

$$t(p) = \alpha * \frac{\beta t(p)}{\beta t(p) + (1-\beta)d(p)} * \sum_{q:q\to p} \frac{t(q)}{outDegree(q)} + (1-\alpha)s(p) \qquad (۱۰.۴)$$

$$d(p) = \alpha' * \frac{(1-\beta)d(p)}{(1-\beta)d(p) + \beta t(p)} * \sum_{q:p\to q} \frac{d(q)}{inDegree(q)} + (1-\alpha')s'(p) \qquad (۱۱.۴)$$

که دو پارامتر α و α' همان عامل میرایی با مقدار ۰/۸۵ می‌باشند. همچنین مقداری که در این پژوهش برای پارامتر β در نظر گرفته شده است، مانند مقاله [۷۲]، برابر با ۰/۵ می‌باشد. $s(p)$ و $s'(p)$ نیز به ترتیب معادل دو پارامتر $s(p)$ و $s'(p)$ در الگوریتم‌های TrustRank و Anti-TrustRank هستند.

- الگوریتم GBR: این الگوریتم مشابه الگوریتم TDR است، با این تفاوت که در زمان انتشار امتیاز از گره i به j، به جای در نظر گرفتن احتمال اعتبار (هرز بودن) گره مقصد، احتمال اعتبار (هرز بودن) گره مبدا در نظر گرفته می‌شود. Liu و همکاران [۷۳] عقیده دارند که وزن انتشار امتیاز اعتماد (عدم اعتماد) صفحات باید با توجه به میزان احتمال اعتماد (عدم اعتماد) صفحه مبدا کنترل شود. بدین معنا که صفحه‌ای که دارای اعتبار زیادی است می‌تواند بخش زیادی از این اعتبار را به صفحات

همسایه انتقال دهد. همچنین یک صفحه هرز که امتیاز اعتبار زیادی ندارد، وزن کمی از آن را می‌تواند به صفحات همسایه دهد و در ازای آن، امتیاز هرز بودن این صفحه با وزن بیشتری به سمت صفحات همسایه انتشار داده می‌شود. روابط ۱۲.۴ و ۱۳.۴ به ترتیب نحوه محاسبه هر یک از امتیازهای اعتبار و هرز بودن صفحات را نشان می‌دهند.

$$g(p) = \alpha \sum_{q:q\to p} \left(\frac{g(q)}{outDegree(q)} * \frac{g(q)}{g(q)+b(q)} \right) + (1-\alpha)s(p) \qquad (۱۲.۴)$$

$$b(p) = \alpha' \sum_{q:p\to q} \left(\frac{b(q)}{inDegree(q)} * \frac{b(q)}{g(q)+b(q)} \right) + (1-\alpha')s'(p) \qquad (۱۳.۴)$$

مقادیر پارامترهای موجود در این دو رابطه نیز، معادل پارامترهای نظیر آن‌ها در روابط ۱۰.۴ و ۱۱.۴ می‌باشد.

همان‌طور که در بخش ۲.۲.۴ توضیح داده شد، برای ارزیابی الگوریتم WorthyRank از معیار ضریب هرز و برای ارزیابی الگوریتم JunkyRank از ضریب اعتماد استفاده می‌نماییم. برای اجرای تمام الگوریتم‌ها از شرط همگرایی ۱۰.۳ با پارامتر $\epsilon = 0.1 * 10^{-9}$ و همچنین پارامتر $\alpha = 0.85$ به عنوان عامل میرایی استفاده شده است.

ارزیابی الگوریتم WorthyRank

برای ارزیابی الگوریتم WorthyRank، آن را با سه روش پیشین TrustRank در رابطه ۸.۴، TRank در رابطه ۱۰.۴، و GRank در رابطه ۱۲.۴ مقایسه می‌نماییم. نتایج هر یک از این روش‌ها در جدول ۸.۴ ارائه شده است.

همان‌طور که مشاهده می‌نماییم، الگوریتم WorthyRank در مقایسه با روش‌های پیشین، در شناسایی وب‌گاه‌های هرز عملکرد بهتری دارد. با استفاده از این روش، میزان خطا در رتبه‌بندی وب‌گاه‌های معتبر به

جدول ۸.۴: نتایج ارزیابی الگوریتم WorthyRank در مقایسه با تعدادی از روش‌های پیشین مربوطه

مجموعه داده‌ای	ضریب هرز			
	TrustRank	TRank	GRank	WorthyRank
WEBSPAM-UK2007	۰/۰۱۵۳	۰/۰۱۴۹	۰/۰۱۳۷	۰/۰۰۸۶
WebSpamChallengeII-CorpusI	۰/۰۰۵۳	۰/۰۰۵۶	۰/۰۰۵۲	۰/۰۰۴۳

مقدار قابل توجهی کاهش یافته است. در ادامه با توجه به نتایج جدول، دلایل ضعف و یا قوت هر یک از این روش‌ها را تحلیل می‌نماییم.

با توجه به جدول ۸.۴، مقدار ضریب هرز در الگوریتم TrustRank در اکثر موارد بیشتر از سایر الگوریتم‌ها می‌باشد. یکی از دلایل پایین بودن کیفیت رتبه‌بندی وب‌گاه‌ها در این الگوریتم، می‌تواند وجود یال‌های جعلی در گراف وب باشد که در این الگوریتم هیچ سیاستی برای مقابله با آن در نظر گرفته نشده است. در این الگوریتم، امتیاز اعتماد هر صفحه بدون توجه به میزان اعتبار پیوند بین دو صفحه، به طور مساوی بین صفحات همسایه خروجی تقسیم می‌شود. هرزنویسان با قرار دادن پیوند صفحات خود درون بخش‌هایی از صفحات معتبر (مانند بخش نظرات کاربران)، این الگوریتم را فریب داده و امتیاز اعتماد صفحات خود را افزایش می‌دهند.

با توجه به نتایج جدول، الگوریتم TRank توانسته است به نسبت الگوریتم TrustRank ضریب هرز کمتری را بدست آورد. همان‌طور که توضیح داده شد، در این الگوریتم امتیاز اعتماد هر صفحه به نسبت احتمال اعتبار صفحه مقصد انتشار داده می‌شود. به عبارت دیگر، وزن یک یال در این الگوریتم، با استفاده از احتمال اعتبار صفحه مقصد مشخص می‌شود. مشکلی که در این روش وجود دارد این است که احتمال اعتباری که برای صفحات محاسبه می‌شود، همچنان با استفاده از ساختار گرافی وب می‌باشد. در این روش، با شروع از یک مجموعه بذر معتبر اولیه، انتشار امتیاز از این مجموعه صفحات به سایر صفحات، در حالی انجام می‌شود که امتیاز اعتماد اولیه صفحات مقصد مشخص نمی‌باشد. بنابراین، صفحات هرزی که پیوند صفحات خود را درون تعداد زیادی از صفحات معتبر قرار می‌دهند، با فریب این الگوریتم می‌توانند رتبه خوبی را در میان سایر صفحات بدست آورند. با مقایسه ضریب هرز این الگوریتم با ضریب هرز الگوریتم

TrustRank برای مجموعه داده‌ای WebSpamChallengeII-CorpusI، مشاهده می‌شود که در مواردی که تعداد این نوع از صفحات هرز در گراف وب زیاد باشد، خطای این الگوریتم از الگوریتم TrustRank نیز بیشتر می‌شود. بنابراین در الگوریتم TRank، تاثیر منفی ناشی از وجود یال‌های جعلی در گراف وب، همچنان در انتشار امتیاز اعتماد از صفحه مبدا به صفحه مقصد وجود دارد.

برای حل این مشکل، الگوریتم GBR معرفی شد که به جای استفاده از احتمال اعتبار صفحه مقصد، از احتمال اعتبار صفحه مبدا استفاده می‌کند. در این الگوریتم با شروع اجرای آن در تکرار اول، احتمال اعتبار مبدا همان امتیاز اعتماد صفحات بذر است که از قبل مشخص شده است. همچنین به این دلیل که این صفحات به صورت دستی برچسب‌گذاری شده‌اند، احتمال اعتبار آن‌ها از دقت زیادی برخوردار است. با توجه به جدول ۸.۴ مشاهده می‌نماییم که ضریب هرز در این الگوریتم کمتر از دو روش TrustRank و TRank می‌باشد. اما همچنان در این الگوریتم خطای ناشی از وجود یال‌های جعلی در گراف وب وجود دارد.

در نهایت، با توجه به جدول ۸.۴، مشاهده می‌نماییم که الگوریتم WorthyRank که در این پژوهش معرفی شده است، نسبت به روش‌های پیشین کمترین ضریب هرز را بدست آورده است. دلیل این امر این است که در این الگوریتم، به جای انتشار امتیاز اعتماد با توجه به احتمال اعتبار صفحه مبدا یا مقصد، از احتمال اعتبار یال‌های گراف استفاده می‌شود. این اعتبار با توجه به سایر گره‌های همسایه و رابطه بین آن‌ها محاسبه می‌شود. در صورتی که در دو روش Trank و Drank، احتمال اعتبار تنها بر اساس یال‌های ورودی به یک صفحه محاسبه می‌شود و یک صفحه هرز می‌تواند به تنهایی با ایجاد یک ساختار پیوندی جعلی بین خود و سایر گره‌های گراف این دو الگوریتم را فریب دهد. در صورتی که در الگوریتم WorthyRank با استفاده از اطلاعاتی که از گره‌های همسایه بدست می‌آید، احتمالی که برای اعتبار هر یال محاسبه می‌شود دقیق‌تر است.

جدول ۹.۴: تاثیر هر یک از بخش‌های الگوریتم WorthyRank در کاهش ضریب هرز

مجموعه داده‌ای	ضریب هرز			
	TrustRank	وزن‌دهی به یال‌ها	بسط دوره‌ای بذر	وزن‌دهی به یال‌ها + بسط دوره‌ای بذر
WEBSPAM-UK2007	۰/۰۱۵۳	۰/۰۰۹۱	۰/۰۱۲۵	۰/۰۰۸۶
WebSpamChallengeII-CorpusI	۰/۰۰۵۳	۰/۰۰۴۳	۰/۰۰۵۴	۰/۰۰۴۳

بررسی میزان کارایی بخش‌های مختلف الگوریتم WorthyRank

در الگوریتم WorthyRank علاوه بر انتخاب بهینه بذر، از دو سیاست خاص، یکی برای وزن‌دهی به یال‌ها و دیگری برای بسط دوره‌ای بذر استفاده شده است. هر یک از این روش‌ها به تنهایی در بهبود الگوریتم‌های انتشار برچسب مانند TrustRank تاثیرات مثبتی دارد که برای بررسی آن‌ها، هر یک را به طور مجزا در الگوریتم TrustRank اعمال می‌نماییم. جدول ۹.۴ میزان تاثیر هر یک از بخش‌های اصلی الگوریتم WorthyRank را در ضریب هرز حاصل از رتبه‌بندی وب‌گاه‌های دو مجموعه داده‌ای WebSpamChallengeII-CorpusI و WEBSPAM-UK2007 نشان می‌دهد.

همان‌طور که در جدول ۹.۴ مشاهده می‌نماییم، تاثیر وزن‌دهی به یال‌های گراف، در بهبود نتایج رتبه‌بندی وب‌گاه‌ها، بیشتر از تاثیر بسط دوره‌ای بذر می‌باشد. همچنین در مورد مجموعه داده‌ای دوم نیز مشاهده می‌کنیم که اعمال بسط دوره‌ای بذر، بدون وزن‌دهی به یال‌ها، سبب افزایش اندک مقدار ضریب هرز شده است. دلیل افزایش خطا می‌تواند این امر باشد که با وجود یال‌های جعلی در گراف وب، تعدادی از وب‌گاه‌های هرز می‌توانند امتیاز اعتماد خود را افزایش داده و به عنوان بذر معتبر جدید به مجموعه وب‌گاه‌های بذر اولیه اضافه شوند. این دسته از وب‌گاه‌ها، وب‌گاه‌های هرزی هستند که پیوند صفحات خود را درون تعداد زیادی از صفحات معتبر قرار می‌دهند. هر اندازه صفحات هرز موجود در گراف، فاصله کمتری با صفحات بذر اولیه داشته باشند، مقدار این خطا بیشتر می‌شود. با توجه به این‌که در مجموعه داده‌ای دوم، متوسط فاصله گره‌ها از همدیگر کوتاه‌تر از متوسط فاصله گره‌ها در گراف مجموعه داده‌ای اول می‌باشد، احتمال ارتباط نزدیک‌تر صفحات هرز دارای یال‌های جعلی، با صفحات بذر معتبر بیشتر است. به همین

جدول ۱۰.۴: نتایج ارزیابی الگوریتم JunkyRank در مقایسه با تعدادی از روش‌های پیشین مربوطه

مجموعه داده‌ای	ضریب اعتماد			
	Anti-TrustRank	DRank	BRank	JunkyRank
WEBSPAM-UK2007	۰/۳۳	۰/۳۶	۰/۳۵	۰/۳۷
WebSpamChallengeII-CorpusI	۰/۷۳	۰/۷۸	۰/۷۴	۰/۸۵

دلیل، مقدار ضریب هرز برای این مجموعه داده‌ای، در روش بسط دوره‌ای بذر افزایش یافته است، که با توجه به نتایج جدول، این مشکل با وزن‌دهی به یال‌های گراف حل می‌شود.

ارزیابی الگوریتم JunkyRank

در این بخش، برای ارزیابی الگوریتم JunkyRank آن را با سه الگوریتم Anti-TrustRank، DRank و BRank مقایسه می‌نماییم. جدول ۱۰.۴ نتایج مربوط به اجرای این سه الگوریتم را بر روی دو مجموعه داده‌ای WebSpamChallengeII-CorpusI و WEBSPAM-UK2007 نشان می‌دهد.

با توجه به جدول ۱۰.۴ مشاهده می‌کنیم که میزان ضریب اعتماد در رتبه‌بندی وب‌گاه‌ها با استفاده از الگوریتم JunkyRank افزایش یافته است. همچنین با توجه به جدول، پس از این الگوریتم به ترتیب الگوریتم‌های DRank و BRank بهترین نتایج را داشته‌اند. دلیل ضعف الگوریتم Anti-TrustRank به نسبت سایر روش‌ها، وجود صفحات هرزی است که پیوند صفحات خود را درون تعدادی از صفحات معتبر قرار می‌دهند. ایجاد چنین پیوندهایی باعث می‌شود که این صفحات، امتیاز هرز بودن خود را به صفحات معتبر انتقال داده و باعث افزایش رتبه هرز بودن صفحات معتبر و در نتیجه کاهش رتبه هرز بودن خود شوند.

در الگوریتم DRank در صورت وجود یال از گره i به گره j، برای انتشار امتیاز از گره j به گره i احتمال هرز بودن گره i در نظر گرفته می‌شود. با توجه به این احتمال، که در صورت معتبر بودن صفحه i مقدار آن کم است، درصد کمی از امتیاز هرز بودن صفحه هرز j به صفحه معتبر i انتقال داده می‌شود. این

مهم در حالی است که در الگوریتم BRank به این دلیل که احتمال هرز بودن صفحه مبدا (j) در نظر گرفته می‌شود، همچنان، وجود پیوندهای جعلی از صفحات معتبر به صفحات هرز در رتبه‌بندی وب‌گاه‌ها تاثیر منفی می‌گذارد. با این حال، این الگوریتم نسبت به روش Anti-TrustRank بهتر عمل می‌کند. دلیل این امر، رفتار خاص وب‌گاه‌های هرز در گراف وب می‌باشد. همان‌طور که در بخش ۳.۲.۳ نیز توضیح داده شد، در یک گراف وب صفحاتی که به یک صفحه هرز ارجاع می‌دهند، با احتمال زیادی خود نیز یک صفحه هرز می‌باشند. در الگوریتم BRank نیز با در نظر گرفتن احتمال هرز بودن صفحه هرز j، امتیاز هرز بیشتری به صفحاتی که به این صفحه ارجاع داده‌اند و با احتمال زیادی هرز هستند، انتقال داده می‌شود.

مشکلی که در هر سه روش پیشین وجود دارد این است که هیچ یک از این الگوریتم‌ها احتمال هرز بودن صفحاتی که یک صفحه هرز به آن‌ها ارجاع داده است را در نظر نمی‌گیرند. در صورتی که در موارد زیادی، مانند صفحات درون دهکده‌های پیوندی، هر صفحه هرز دارای تعداد زیادی ارجاع به سایر صفحات هرز می‌باشد. با استفاده از این روش، هرزنویسان می‌توانند رتبه صفحات خود را افزایش دهند. در الگوریتم JunkyRank برای مقابله با این نوع از صفحات هرز، امتیاز هرز بودن یک صفحه علاوه بر انتشار در خلاف جهت یال‌های گراف، به سمت صفحات همسایه خروجی نیز انتشار داده می‌شود. همان‌طور که نتایج جدول ۱۰.۴ نیز نشان می‌دهد، با استفاده از این روش می‌توان کیفیت رتبه‌بندی وب‌گاه‌ها را بهبود داد.

۳.۳.۴ ارزیابی روش ترکیبی محتوایی و پیوندی در شناسایی هرز وب

در این بخش، نتایج مربوط به رتبه‌بندی وب‌گاه‌ها با استفاده از الگوریتم CLCRank ارائه می‌شود. برای اجرای این الگوریتم که از ترکیب اطلاعات محتوایی و پیوندی استفاده می‌کند، از دو مجموعه داده‌ای استاندارد WebSpamChallengeII-CorpusI و WEBSPAM-UK2007 استفاده شده است. همچنین، مقایسه نتایج بر اساس دو معیار ضریب هرز و ضریب اعتماد صورت می‌گیرد. برای ارزیابی الگوریتم CLCRank، نتایج آن را با نتایج حاصل از بهترین روش ترکیبی معرفی شده در مقاله [۱۱۱] مقایسه می‌نماییم. بدین منظور، ابتدا به توضیح مختصری پیرامون این روش ترکیبی به عنوان یک روش پایه می‌پردازیم.

معرفی روش پایه

در این روش که در سال ۲۰۱۲ توسط Ortega و همکاران [۱۱۱] معرفی شده است، از سه روش برای ترکیب اطلاعات محتوایی با الگوریتم انتشار برچسب استفاده شده است که در این میان، روش CS-NS بیشترین دقت را در رتبه‌بندی وب‌گاه‌های معتبر داشته است. در این روش، ابتدا برای هر صفحه مقدار برخی از ویژگی‌های محتوایی که با احتمال هرز بودن وب‌گاه‌ها رابطه مستقیم دارند، محاسبه شده و بردار F_p برای هر صفحه P، با استفاده از مقادیر این ویژگی‌ها مقداردهی می‌شود. سپس امتیاز هرز بودن هر صفحه P با استفاده از رابطه ۱۴.۴ محاسبه می‌شود.

$$spaminess(P) = \sqrt{\sum_{h \in F_p} h^2} \tag{۱۴.۴}$$

Ortega و همکاران [۱۱۱] با بررسی هزینه محاسباتی تعدادی از ویژگی‌های محتوایی، در نهایت دو ویژگی محتوایی «متوسط طول کلمات» و «درصد فشرده‌سازی صفحه» را به دلیل هزینه محاسباتی کم آن انتخاب نمودند. با داشتن احتمال هرز بودن صفحات، N^- صفحه با بیشترین امتیاز هرز، به عنوان مجموعه صفحات بذر هرز (S^-) و N^+ صفحه با کمترین امتیاز هرز بودن، به عنوان مجموعه صفحات بذر معتبر (S^+) انتخاب می‌شوند. سپس وزن هر صفحه بذر معتبر و هرز i به ترتیب با استفاده از روابط ۱۵.۴ و ۱۶.۴ محاسبه می‌شود:

$$e_i^+ = \begin{cases} \frac{spaminess(i)}{\sum_{j \in S^+} spaminess(j)}, & \text{if } i \in S^+ \\ 0, & \text{otherwise} \end{cases} \tag{۱۵.۴}$$

$$e_i^- = \begin{cases} \frac{spaminess(i)}{\sum_{j \in S^-} spaminess(j)}, & \text{if } i \in S^- \\ 0, & \text{otherwise} \end{cases} \tag{۱۶.۴}$$

پس از آن، امتیازهای اعتبار و هرز محتوایی بدست آمده از روابط ۱۵.۴ و ۱۶.۴ به ترتیب با استفاده از روابط ۱۷.۴ و ۱۸.۴ در کل گراف انتشار می‌یابد.

$$PR^{+}(p) = (1-\alpha){e_i}^{+} + \alpha \sum_{q:q\to p} \frac{PR^{+}(q)}{outDegree(q)} \quad (۱۷.۴)$$

$$PR^{-}(p) = (1-\alpha){e_i}^{-} + \alpha \sum_{q:p\to q} \frac{PR^{-}(q)}{inDegree(q)} \quad (۱۸.۴)$$

در نهایت، امتیاز نهایی هر صفحه به صورت زیر محاسبه می‌شود:

$$score(p) = PR^{+}(p) - PR^{-}(p) \quad (۱۹.۴)$$

ارزیابی روش CLCRank

برای اجرای الگوریتم‌های این بخش، از پارامتر $\alpha = 0/85$ به عنوان عامل میرایی و شرط همگرایی ۱۰.۳ با پارامتر $\epsilon = 0.1 * 10^{-9}$ استفاده نمودیم.

نتایج مربوط به رتبه‌بندی وب‌گاه‌ها با استفاده از دو روش CLCRank و CS-NS در جدول ۱۱.۴ ارائه شده است. این نتایج نشان می‌دهد که میزان خطا در رتبه‌بندی وب‌گاه‌های معتبر با استفاده از الگوریتم CLCRank به میزان قابل توجهی کمتر از الگوریتم CS-NS است. لازم به ذکر است که، همان‌طور که در بخش ۱.۴ نیز توضیح داده شد، تنها اطلاعاتی که از محتوای وب‌گاه‌های درون مجموعه داده‌ای WebSpamChallengeII-CorpusI در اختیار داریم، وزن TF-IDF کلمات درون هر وب‌گاه است. با استفاده از این وزن می‌توان درصد فشرده‌سازی هر وب‌گاه را محاسبه کرد. برای محاسبه این ویژگی، ابتدا مقدار TF هر کلمه را حساب کرده، سپس مجموع TF کلمات درون هر وب‌گاه را بر تعداد کلمات یکتای آن وب‌گاه تقسیم می‌نماییم. همچنین، به دلیل محدودیت دسترسی به اطلاعات محتوایی، به جای استفاده از ویژگی

جدول ۱۱.۴: مقایسه ضریب هرز در روش CLCRank و روش پایه CS-NS

مجموعه داده‌ای	ضریب هرز	
	CS-NS	CLCRank
WEBSPAM-UK2007	۰/۰۳۳۸	۰/۰۰۸۴
WebSpamChallengeII-CorpusI	۰/۰۰۶۷	۰/۰۰۴۲

«متوسط طول کلمات»، از ویژگی «تعداد کلمات درون وب‌گاه» استفاده کردیم. همان‌طور که در [۸] نیز نشان داده شده است، مقدار این ویژگی، مانند ویژگی «متوسط طول کلمات»، با احتمال هرز بودن وب‌گاه‌ها رابطه مستقیم دارد. همچنین هزینه محاسباتی این ویژگی نیز از ویژگی «متوسط طول کلمات» کمتر می‌باشد. بنابراین جایگزینی این ویژگی با ویژگی «متوسط طول کلمات»، در کارایی الگوریتم تاثیر منفی ندارد.

در ادامه، برای ارزیابی میزان کارایی الگوریتم CLCRank در محاسبه امتیاز هرز بودن هر وب‌گاه، مقدار ضریب اعتماد این الگوریتم را با ضریب اعتماد بدست آمده از الگوریتم پایه CS-NS مقایسه می‌نماییم. برای محاسبه ضریب اعتماد در الگوریتم CS-NS، برای هر وب‌گاه امتیازی را در نظر می‌گیریم که با استفاده از رابطه ۲۰.۴ بدست آمده است.

$$score(p) = PR^{-}(p) - PR^{+}(p) \qquad (۲۰.۴)$$

جدول ۱۲.۴ نتایج حاصل از رتبه‌بندی وب‌گاه‌ها با استفاده از دو الگوریتم CLCRank و CS-NS را نشان می‌دهد.

جدول ۱۲.۴: مقایسه ضریب اعتماد در روش CLCRank و روش پایه CS-NS

مجموعه داده‌ای	ضریب اعتماد	
	CS-NS	**CLCRank**
WEBSPAM-UK2007	۰/۳۳	۰/۴۰
WebSpamChallengeII-CorpusI	۰/۷۴	۰/۸۹

بررسی نتایج دو جدول ۱۱.۴ و ۱۲.۴ نشان می‌دهد که استفاده از روش ترکیبی CLCRank کیفیت رتبه‌بندی وب‌گاه‌ها را به نسبت روش پایه CS-NS به میزان قابل توجهی بهبود می‌دهد. همچنین مقایسه نتایج این بخش با نتایج بخش ۲.۳.۴، نکاتی را پیرامون کارایی الگوریتم‌های ترکیبی نسبت به الگوریتم‌های پیوندی آشکار می‌سازد که در ادامه به شرح مهم‌ترین آن‌ها می‌پردازیم.

در مرحله اول، مقایسه نتایج مربوط به الگوریتم‌های WorthyRank و JunkyRank با نتایج حاصل از الگوریتم CLCRank نشان می‌دهد که ترکیب اطلاعات محتوایی با روش‌های پیوندی، تاثیر قابل توجهی در بهبود کیفیت رتبه‌بندی وب‌گاه‌ها دارد. در الگوریتم CLCRank با استفاده از احتمال اعتبار (عدم اعتبار) محتوایی وب‌گاه‌ها، می‌توان میزان انتشار امتیاز اعتماد (عدم اعتماد) وب‌گاه‌ها را در طول هر مسیر کنترل کرد. با استفاده از این روش، از خطای ناشی از وجود یال‌های جعلی در گراف وب کاسته می‌شود.

از طرف دیگر، با مقایسه روش پایه ترکیبی CS-NS با دو روش پایه مبتنی بر پیوند TrustRank و Anti-TrustRank، مشاهده می‌نماییم که استفاده از روش ترکیبی CS-NS، نه‌تنها تاثیر چندانی در بهبود معیار ضریب اعتماد ندارد، بلکه ضریب هرز را نیز که نشان‌گر میزان خطا در رتبه‌بندی وب‌گاه‌ها می‌باشد، افزایش داده است. یکی از دلایل مهم این امر، سیاست نامناسب انتخاب وب‌گاه‌های بذر در این الگوریتم می‌باشد. در این الگوریتم، انتخاب مجموعه وب‌گاه‌های بذر بر اساس ویژگی‌های محتوایی انجام می‌شود، در صورتی که کاربرد اصلی این وب‌گاه‌ها، در انتشار برچسب می‌باشد و بهتر است که بر اساس معیارهای مبتنی بر پیوند انتخاب شوند. بنابراین، اگرچه ممکن است یک وب‌گاه، با توجه به امتیاز محتوایی به عنوان یک وب‌گاه هرز شناسایی شود، اما این وب‌گاه لزوما یک بذر مناسب برای الگوریتم‌های انتشار برچسب نمی‌باشد. به عنوان مثال، این احتمال وجود دارد که این وب‌گاه مربوط به گره‌ای در گراف وب باشد که با سایر گره‌ها ارتباط زیادی ندارد. بدیهی است که چنین وب‌گاهی نمی‌تواند امتیاز خود را به طور بهینه در سراسر گراف وب انتشار دهد. به همین دلیل است که در الگوریتم CLCRank، برای انتخاب بذر از روش‌های پیوندی مانند PageRank استفاده می‌نماییم.

مشکل دیگر این الگوریتم، نحوه ترکیب امتیاز محتوایی با روش پیوندی می‌باشد. در این الگوریتم، تنها از امتیاز محتوایی مربوط به وب‌گاه‌های بذر استفاده می‌شود و فقط این امتیازها در گراف انتشار داده

می‌شوند. در صورتی که الگوریتم CLCRank، از امتیاز محتوایی تمام وب‌گاه‌های درون گراف استفاده می‌کند.

یکی از مشکلات اساسی در الگوریتم CS-NS، وزن‌دهی وب‌گاه‌های بذر معتبر بر اساس امتیاز هرز بودن آن‌ها است. با توجه به رابطه ۱۵.۴ مشاهده می‌نماییم که برای مقداردهی به وب‌گاه‌های بذر معتبر نیز مانند وب‌گاه‌های هرز، از امتیاز هرز استفاده شده است؛ به عبارت دیگر، با توجه به این رابطه، هر میزان مقدار هرز بودن یک وب‌گاه معتبر بیشتر باشد، وزن اولیه آن وب‌گاه نیز بیشتر است. در صورتی که در مورد وب‌گاه‌های معتبر، عکس این امر صدق می‌کند. دلیل افزایش قابل توجه ضریب هرز در این الگوریتم نسبت به روش‌های قبلی نیز همین مساله می‌باشد.

فصل ۵

جمع‌بندی و نکته‌های پایانی

در این پژوهش، سعی کردیم تاثیر انواع ویژگی‌های محتوایی را در شناسایی وب‌گاه‌های هرز فارسی بررسی کرده و یک روش با کارایی بالا برای شناسایی این نوع از وب‌گاه‌ها ارائه دهیم. همچنین، با توجه به خصوصیات خاص وب‌گاه‌های هرز، دو روش مبتنی بر پیوند برای رتبه‌بندی وب‌گاه‌ها ارائه دادیم. در نهایت برای بهبود کیفیت رتبه‌بندی، روشی را معرفی کردیم که از ویژگی‌های محتوایی نیز در کنار ویژگی‌های پیوندی استفاده می‌کند. در این فصل، ابتدا مرور مختصری بر دستاوردهای این پژوهش داشته و در ادامه، به منظور بهبود و گسترش این پژوهش، پیشنهادهایی را برای کارهای آینده ارائه می‌دهیم.

۱.۵ دستاوردهای پژوهش

با توجه به نبود مطالعات گسترده در زمینه شناسایی و مقابله با وب‌گاه‌های هرز فارسی، در این پژوهش ابتدا به بررسی کارایی تعداد زیادی از ویژگی‌های محتوایی پیشین در شناسایی وب‌گاه‌های فارسی پرداختیم. بررسی این ویژگی‌ها بر روی وب‌گاه‌های فارسی نشان داده است که بسیاری از وب‌گاه‌های هرز فارسی دارای خصوصیاتی هستند که با استفاده از ویژگی‌های محتوایی مربوط به خصوصیات ساختاری وب‌گاه‌ها، قابل شناسایی نمی‌باشند. همچنین نشان دادیم که تعدادی از ویژگی‌های محتوایی در رده‌بندی وب‌گاه‌های

فارسی تاثیر منفی داشته و باعث کاهش دقت رده‌بندی می‌شوند. برای بهبود کارایی رده‌بندی این نوع از وب‌گاه‌ها، تعدادی ویژگی محتوایی جدید معرفی کردیم. هر چند با استفاده از این ویژگی‌ها، کیفیت شناسایی وب‌گاه‌های هرز بهبود پیدا کرد، اما همچنان امکان شناسایی انواع خاصی از وب‌گاه‌های فارسی وجود نداشت. بنابراین، یک سامانه شناساگر هرز وب به نام PSD-SYS را ارائه دادیم که از مدل جدیدی به نام BOSW برای شناسایی وب‌گاه‌های هرز فارسی استفاده می‌کند. این سامانه، با استفاده از مدل BOSW، که بر مبنای کلمات رایج در وب‌گاه‌های هرز است، توانسته است درصد زیادی از وب‌گاه‌های هرز فارسی را با دقت خوبی تشخیص دهد. از جمله مزایای این سامانه این است که علاوه بر عملکرد خوب در رده‌بندی وب‌گاه‌ها، به نسبت سایر روش‌های محتوایی هزینه زمانی و محاسباتی کمتری دارد.

در ادامه، در این پژوهش به بررسی روش‌های مبتنی بر پیوند پرداخته و با یافتن نقاط ضعف الگوریتم‌هایی که تاکنون مطرح شده‌اند، دو روش مبتنی بر پیوند معرفی کردیم که مشکلات الگوریتم‌های پیشین را ندارند. در الگوریتم WorthyRank، با استفاده از وزن‌دهی به یال‌های گراف و مشخص کردن میزان اعتبار آن‌ها، خطاهای ناشی از وجود پیوندهای جعلی را در گراف وب کاهش دادیم. مزیت دیگری که این روش نسبت به روش‌های پیشین دارد، نیمه‌سرپرست بودن این الگوریتم است که این امکان را می‌دهد که با داشتن تعداد محدودی از وب‌گاه‌های برچسب خورده، در هر تکرار الگوریتم به صورت خودکار تعدادی وب‌گاه جدید به مجموعه وب‌گاه‌های بذر اولیه اضافه شود. این روش برای وظیفه شناسایی هرز وب، که برچسب‌زنی وب‌گاه‌ها در آن کاری زمان‌بر و پرهزینه می‌باشد، بسیار مناسب است.

در الگوریتم دیگری به نام JunkyRank، نشان دادیم که امتیاز هرز بودن یک وب‌گاه، علاوه بر انتشار به صورت پس‌رو در گراف، باید به صورت پیش‌رو نیز انتشار داده شود. دلیل این امر، وجود صفحات هرزی است که بخشی از یک دهکده پیوندی می‌باشند. وجود دهکده پیوندی در میان وب‌گاه‌ها امری رایج است که توسط هرزنویسان و با هدف افزایش رتبه وب‌گاه‌های هرز ایجاد می‌شود. با استفاده از آزمایش‌هایی که بر روی این الگوریتم صورت گرفت، نشان دادیم که استفاده از این روش در مقایسه با روش‌هایی که امتیاز هرز را فقط به صورت پس‌رو در گراف وب انتشار می‌دهند، نتایج بهتری را در رده‌بندی وب‌گاه‌ها بدست می‌آورد.

در نهایت نیز یک الگوریتم به نام CLCRank را پیشنهاد دادیم که در آن، علاوه بر استفاده از اطلاعات ساختاری گراف وب، از امتیازاتی که با استفاده از رده‌بند محتوایی برای وب‌گاه‌ها محاسبه شده است نیز استفاده می‌شود. با استفاده از این روش نشان دادیم که برای شناسایی بهتر وب‌گاه‌های هرز، به هر دو نوع اطلاعات محتوایی و پیوندی آن‌ها نیاز داریم. این امر به این دلیل است که هرزنویسان برای فریب موتورهای جست‌وجو، از روش‌های متفاوتی برای افزایش رتبه وب‌گاه‌های هرز استفاده می‌کنند. برای مثال، در حالی‌که در برخی از وب‌گاه‌های هرز، فقط از روش‌های محتوایی برای افزایش رتبه استفاده شده است؛ برخی دیگر، تنها از روش‌های پیوندی استفاده می‌کنند. بسیاری از وب‌گاه‌های هرز نیز از ترکیب روش‌های موجود، برای افزایش رتبه خود در میان نتایج موتورهای جست‌وجو استفاده می‌کنند.

یکی دیگر از دستاوردهای مهم این پژوهش، ساخت پیکره‌ای از وب‌گاه‌های فارسی برچسب‌خورده به نام PersianWebSpam-2013 می‌باشد. با توجه به نبود پژوهشی قابل توجه در زمینه شناسایی و مقابله با وب‌گاه‌های هرز فارسی، مجموعه داده‌ای مناسبی در این زمینه در دسترس نبود. بنابراین برای انجام این پژوهش، ابتدا به جمع‌آوری و ساخت یک مجموعه‌داده‌ای از وب‌گاه‌های فارسی، که شامل ۳۰۰ وب‌گاه هرز و ۱۰۵۰ وب‌گاه معتبر می‌باشد، پرداخته شد.

۲.۵ کارهای آینده

در این بخش، تعدادی از کارهای آتی که به منظور گسترش و بهبود هر یک از روش‌های معرفی شده در این پژوهش می‌توان انجام داد را معرفی می‌نماییم.

از جمله مهم‌ترین کارهایی که می‌توان در راستای بهبود کیفیت مطالعات انجام شده در زمینه شناسایی هرز وب فارسی انجام داد، گسترش مجموعه داده‌ای PersianWebSpam-2013 و ایجاد یک مجموعه داده‌ای جامع‌تر، از وب‌گاه‌های هرز و معتبر فارسی می‌باشد. با توجه به این‌که با گذشت زمان انواع جدیدی از وب‌گاه‌های هرز ایجاد می‌شود، با استفاده از مجموعه‌های داده‌ای پیشین نمی‌توان ارزیابی درستی از روش‌های شناسایی هرز وب داشت. در مجموعه داده‌ای جدید، علاوه بر برچسب هرز و معتبر، می‌توان

برچسب خط مرزی را نیز در نظر گرفت. این برچسب به صفحاتی داده می‌شود که نمی‌توان در مورد هرز بودن یا نبودن آن‌ها قضاوت کرد. با توجه به میزان قابل توجه این صفحات در وب، اضافه کردن برچسب مربوط به آن‌ها در رده‌بندی وب‌گاه‌ها، می‌تواند تاثیر قابل توجهی در بهبود کیفیت رده‌بندی وب‌گاه‌ها داشته باشد.

برای داشتن یک نمونه مناسب از گراف وب، امکان جدا کردن صفحات وب بر اساس زبان آن‌ها وجود ندارد و لازم است که در مرحله خزش، تمام صفحات و پیوندهای بین آن‌ها در نظر گرفته شود. سپس در مراحل بعد، می‌توان طی انجام فرآیند پیش‌پردازش داده‌ها، محدودیت‌هایی را با توجه به زبان صفحات اعمال کرد. همچنین در مرحله خزش نیز می‌توان با انتخاب صفحات بذر به زبان موردنظر، صفحات درون گراف را به سمت این نوع از صفحات متمایل کرد. با استفاده از این روش، می‌توان یک مجموعه‌داده‌ای مناسب شامل صفحات فارسی ایجاد کرد. با توجه به این‌که به دلیل محدودیت زمانی انجام این پژوهش و همچنین هزینه زمانی بالای برچسب‌گذاری صفحات، امکان ساخت و تکمیل مجموعه داده‌ای با حجم زیاد وجود نداشت، امید است که بتوان برای پژوهش‌های بعدی یک مجموعه داده‌ای کامل شامل درصد زیادی از صفحات فارسی را تهیه کرده و در دسترس پژوهشگران قرار بدهیم.

یکی از مشکلاتی که در اکثر روش‌های محتوایی مطرح شده وجود دارد، ضرورت وجود مجموعه‌ای از صفحات برچسب‌خورده، به عنوان داده‌های آموزش می‌باشد. همان‌طور که می‌دانیم برچسب‌گذاری صفحات وب، کاری زمان‌بر و پرهزینه می‌باشد. همچنین با توجه به سرعت زیاد تغییر محتوای صفحات، به‌روز رسانی داده‌های آموزش امری مهم می‌باشد. بنابراین با ارائه روش‌های بدون‌سرپرست، می‌توان هزینه ناشی از ساخت داده‌های آموزش را کاهش داد. علاوه بر این امر، با استفاده از روش‌های بدون‌سرپرست، خطای ناشی از سوگیری نتایج به سمت داده‌های آموزش حذف می‌شود.

همان‌طور که در بخش ۳.۲ توضیح داده شد، یکی از روش‌های شناسایی هرز وب، استفاده از بازخورد[1] های کاربران و اطلاعات مربوط به تاریخچه رفتار آن‌ها نسبت به وب‌گاه‌های مختلف می‌باشد. برای مثال، با دانستن تعداد دفعاتی که کاربران بر روی پیوند یک وب‌گاه تلیک می‌کنند، و یا مدت زمانی که

[1] feedback

در آن وب‌گاه می‌مانند، می‌توان تخمینی از میزان مفید بودن محتوای آن وب‌گاه داشت. ترکیب این اطلاعات با روش‌هایی که در این پژوهش مورد بررسی قرار گرفته‌اند، می‌تواند در بهبود کارایی روش‌های شناسایی هرز وب تاثیر مثبت داشته باشد. همچنین، انواعی از صفحات هرز هستند که بخشی از محتوای آن‌ها، مانند عنوان صفحه و یا بخش ابربرچسب کلیدواژه‌ها و توضیحات، به صورت پویا ایجاد می‌شود. این صفحات، به صورت برخط و پویا، در زمان وارد کردن یک پرس‌وجو توسط کاربران، کلمات آن پرس‌وجو را درون صفحه خود تکرار می‌کنند. ارائه روش‌هایی برای شناسایی این نوع از صفحات هرز، از اهمیت ویژه‌ای برخوردار است.

یکی از سیاست‌هایی که در الگوریتم WorthyRank برای بهبود کیفیت رتبه‌بندی وب‌گاه‌ها استفاده شد، محاسبه میزان اعتبار یال‌های درون گراف وب و وزن‌دهی آن‌ها بر اساس این میزان اعتبار می‌باشد. در این پژوهش برای وزن‌دهی به یال‌ها از ضریب جاکارد استفاده شده است. برای محاسبه این ضریب به ازای هر زوج گره، فقط همسایه‌های مستقیم دو گره و رابطه بین آن‌ها بررسی می‌شود. با توجه به گستردگی رابطه بین گره‌ها در گراف وب، استفاده از اطلاعات همسایه‌های سطح دو و بیشتر، می‌تواند باعث بهبود دقت وزن‌دهی یال‌های گراف شود.

مراجع

[1] E. Convey, "Porn sneaks way back on web," *The Boston Herald*, p.028, 1996.

[2] Z. Gyongyi and H. Garcia-Molina, "Web spam taxonomy," in *First international workshop on adversarial information retrieval on the web (AIRWeb 2005)*, 2005.

[3] A. Perkins, "The classification of search engine spam," 2001.

[4] M. R. Henzinger, R. Motwani, and C. Silverstein, "Challenges in web search engines," in *ACM SIGIR Forum*, vol.36, pp.11–22, ACM, 2002.

[5] D. Fetterly, "Adversarial information retrieval: The manipulation of web content," *ACM Computing Reviews*, 2007.

[6] R. Jennings, "The global economic impact of spam," *Ferris Research*, 2005.

[7] R. Jennings, "Cost of spam is flattening–our 2009 predictions," *Ferris Research*, 2009.

[8] A. Ntoulas, M. Najork, M. Manasse, and D. Fetterly, "Detecting spam web pages through content analysis," in *Proceedings of the 15th international conference on World Wide Web*, pp.83–92, ACM, 2006.

[9] C. Silverstein, H. Marais, M. Henzinger, and M. Moricz, "Analysis of a very large web search engine query log," in *ACm SIGIR Forum*, vol.33, pp.6–12, ACM, 1999.

[10] B. J. Jansen and A. Spink, "An analysis of web documents retrieved and viewed.," in *International Conference on Internet Computing*, pp.65–69, Citeseer, 2003.

[11] G. Salton, A. Wong, and C.-S. Yang, "A vector space model for automatic indexing," *Communications of the ACM*, vol.18, no.11, pp.613–620, 1975.

[12] S. Robertson, H. Zaragoza, and M. Taylor, "Simple bm25 extension to multiple weighted fields," in *Proceedings of the thirteenth ACM international conference on Information and knowledge management*, pp.42–49, ACM, 2004.

[13] C. Zhai, "Statistical language models for information retrieval," *Synthesis Lectures on Human Language Technologies*, vol.1, no.1, pp.1–141, 2008.

[14] "Usage of content languages for websites," `http://w3techs.com/technologies/overview/content_language/all`, accessed August, 2014.

[15] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web.," 1999.

[16] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment," *Journal of the ACM (JACM)*, vol.46, no.5, pp.604–632, 1999.

[17] Z. Gyöngyi, H. Garcia-Molina, and J. Pedersen, "Combating web spam with trustrank," in *Proceedings of the Thirtieth international conference on Very large data bases-Volume 30*, pp.576–587, VLDB Endowment, 2004.

[18] S. Adali, T. Liu, and M. Magdon-Ismail, "Optimal link bombs are uncoordinated.," in *AIRWeb*, pp.58–69, 2005.

[19] Z. Gyöngyi and H. Garcia-Molina, "Link spam alliances," in *Proceedings of the 31st international conference on Very large data bases*, pp.517–528, VLDB Endowment, 2005.

[20] B. D. Davison, "Recognizing nepotistic links on the web," *Artificial Intelligence for Web Search*, pp.23–28, 2000.

[21] D. Fetterly, M. Manasse, and M. Najork, "Spam, damn spam, and statistics: Using statistical analysis to locate spam web pages," in *Proceedings of the 7th International Workshop on the Web and Databases: colocated with ACM SIGMOD/PODS 2004*, pp.1–6, ACM, 2004.

[22] D. Fetterly, M. Manasse, M. Najork, and J. Wiener, "A large-scale study of the evolution of web pages," in *Proceedings of the 12th international conference on World Wide Web*, pp.669–678, ACM, 2003.

[23] D. Fetterly, M. Manasse, and M. Najork, "Detecting phrase-level duplication on the world wide web," in *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pp.170–177, ACM, 2005.

[24] A. Z. Broder, S. C. Glassman, M. S. Manasse, and G. Zweig, "Syntactic clustering of the web," *Computer Networks and ISDN Systems*, vol.29, no.8, pp.1157–1166, 1997.

[25] A. Z. Broder, "Some applications of rabin's fingerprinting method," in *Sequences II*, pp.143–152, Springer, 1993.

[26] M. O. Rabin. *Fingerprinting by random polynomials*. Center for Research in Computing Techn., Aiken Computation Laboratory, Univ., 1981.

[27] L. Breiman, "Bagging predictors," *Machine learning*, vol.24, no.2, pp.123–140, 1996.

[28] Y. Freund and R. E. Schapire, "A desicion-theoretic generalization of on-line learning and an application to boosting," in *Computational learning theory*, pp.23–37, Springer, 1995.

[29] V. M. Prieto, M. Álvarez, and F. Cacheda, "Saad, a content based web spam analyzer and detector," *Journal of Systems and Software*, vol.86, no.11, pp.2906–2918, 2013.

[30] M. Sydow, J. Piskorski, D. Weiss, and C. Castillo, "Application of machine learning in combating web spam," 2007.

[31] J. Piskorski, M. Sydow, and D. Weiss, "Exploring linguistic features for web spam detection: A preliminary study," in *Proceedings of the 4th international workshop on Adversarial information retrieval on the web*, pp.25–28, ACM, 2008.

[32] J. Martinez-Romo and L. Araujo, "Web spam identification through language model analysis," in *Proceedings of the 5th international workshop on adversarial information retrieval on the web*, pp.21–28, ACM, 2009.

[33] G. Mishne, D. Carmel, and R. Lempel, "Blocking blog spam with language model disagreement.," in *AIRWeb*, vol.5, pp.1–6, 2005.

[34] T. Urvoy, T. Lavergne, and P. Filoche, "Tracking web spam with hidden style similarity.," in *AIRWeb*, pp.25–31, 2006.

[35] I. Bíró, J. Szabó, and A. A. Benczúr, "Latent dirichlet allocation in web spam filtering," in *Proceedings of the 4th international workshop on Adversarial information retrieval on the web*, pp.29–32, ACM, 2008.

[36] A. Pavlov and B. V. Dobrov, "Detecting content spam on the web through text diversity analysis.," in *SYRCoDIS*, pp.11–18, 2011.

[37] C. Dong and B. Zhou, "Effectively detecting content spam on the web using topical diversity measures," in *Proceedings of the The 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology-Volume 01*, pp.266–273, IEEE Computer Society, 2012.

[38] Y. Suhara, H. Toda, S. Nishioka, and S. Susaki, "Automatically generated spam detection based on sentence-level topic information," in *Proceedings of the 22nd international conference on World Wide Web companion*, pp.1157–1160, International World Wide Web Conferences Steering Committee, 2013.

[39] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *the Journal of machine Learning research*, vol.3, pp.993–1022, 2003.

[40] M. Riedl and C. Biemann, "Sweeping through the topic space: bad luck? roll again!," in *Proceedings of the Joint Workshop on Unsupervised and Semi-Supervised Learning in NLP*, pp.19–27, Association for Computational Linguistics, 2012.

[41] H. A. Wahsheh and M. N. Al-Kabi, "Detecting arabic web spam," in *The 5th International Conference on Information Technology, ICIT*, vol.2011, pp.1–8, 2011.

[42] R. Jaramh, T. Saleh, S. Khattab, and I. Farag, "Detecting arabic spam web pages using content analysis," *International Journal of Reviews in Computing*, vol.6, pp.1–8, 2011.

[43] M. Al-Kabi, H. Wahsheh, A. AlEroud, and I. Alsmadi, "Combating arabic web spam using content analysis," in *2011 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT)*, pp.1–4, 2011.

[44] M. Al-Kabi, H. Wahsheh, I. Alsmadi, E. Al-Shawakfa, A. Wahbeh, and A. Al-Hmoud, "Content-based analysis to detect arabic web spam," *Journal of Information Science*, vol.38, no.3, pp.284–296, 2012.

[45] M. N. Al-Kabi, H. A. Wahsheh, and I. M. Alsmadi, "Olawsds: An online arabic web spam detection system," *International Journal of Advanced Computer Science & Applications*, vol.5, no.2, 2014.

[46] E. Amitay, D. Carmel, A. Darlow, R. Lempel, and A. Soffer, "The connectivity sonar: detecting site functionality by structural patterns," in *Proceedings of the fourteenth ACM conference on Hypertext and hypermedia*, pp.38–47, ACM, 2003.

[47] L. Becchetti, C. Castillo, D. Donato, S. Leonardi, and R. A. Baeza-Yates, "Link-based characterization and detection of web spam.," in *AIRWeb*, pp.1–8, 2006.

[48] I. Drost and T. Scheffer, "Thwarting the nigritude ultramarine: Learning to identify link spam," in *Machine Learning: ECML 2005*, pp.96–107, Springer, 2005.

[49] B. Wu and B. D. Davison, "Identifying link farm spam pages," in *Special interest tracks and posters of the 14th international conference on World Wide Web*, pp.820–829, ACM, 2005.

[50] B. Wu and B. D. Davison, "Undue influence: Eliminating the impact of link plagiarism on web search rankings," in *Proceedings of the 2006 ACM symposium on Applied computing*, pp.1099–1104, ACM, 2006.

[51] R. Lempel and S. Moran, "The stochastic approach for link-structure analysis (salsa) and the tkc effect," *Computer Networks*, vol.33, no.1, pp.387–401, 2000.

[52] J. Abernethy, O. Chapelle, and C. Castillo, "Graph regularization methods for web spam detection," *Machine Learning*, vol.81, no.2, pp.207–225, 2010.

[53] Z. Cheng, B. Gao, C. Sun, Y. Jiang, and T.-Y. Liu, "Let web spammers expose themselves," in *Proceedings of the fourth ACM international conference on Web search and data mining*, pp.525–534, ACM, 2011.

[54] D. Zhou, C. J. Burges, and T. Tao, "Transductive link spam detection," in *Proceedings of the 3rd international workshop on Adversarial information retrieval on the web*, pp.21–28, ACM, 2007.

[55] C. Castillo, D. Donato, A. Gionis, V. Murdock, and F. Silvestri, "Know your neighbors: Web spam detection using the web topology," in *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pp.423–430, ACM, 2007.

[56] J. Caverlee and L. Liu, "Countering web spam with credibility-based link analysis," in *Proceedings of the twenty-sixth annual ACM symposium on Principles of distributed computing*, pp.157–166, ACM, 2007.

[57] A. Joshi, R. Kumar, B. Reed, and A. Tomkins, "Anchor-based proximity measures," in *Proceedings of the 16th international conference on World Wide Web*, pp.1131–1132, ACM, 2007.

[58] Q. Gan and T. Suel, "Improving web spam classifiers using link structure," in *Proceedings of the 3rd international workshop on Adversarial information retrieval on the web*, pp.17–20, ACM, 2007.

[59] R. Baeza-Yates, P. Boldi, and C. Castillo, "Generalizing pagerank: Damping functions for link-based ranking algorithms," in *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pp.308–315, ACM, 2006.

[60] V. Krishnan and R. Raj, "Web spam detection with anti-trust rank.," in *AIRWeb*, vol.6, pp.37–40, 2006.

[61] B. Wu, V. Goel, and B. D. Davison, "Propagating trust and distrust to demote web spam.," *MTW*, vol.190, 2006.

[62] L. Becchetti, C. Castillo, D. Donato, S. Leonardi, and R. Baeza-Yates, "Using rank propagation and probabilistic counting for link-based spam detection," in *Proc. of WebKDD*, vol.6, 2006.

[63] G.-G. Geng, Q. Li, and X. Zhang, "Link based small sample learning for web spam detection," in *Proceedings of the 18th international conference on World wide web*, pp.1185–1186, ACM, 2009.

[64] B. Wu, V. Goel, and B. D. Davison, "Topical trustrank: Using topicality to combat web spam," in *Proceedings of the 15th international conference on World Wide Web*, pp.63–72, ACM, 2006.

[65] Q. Chen, S.-N. Yu, and S. Cheng, "Link variable trustrank for fighting web spam," in *Computer Science and Software Engineering, 2008 International Conference on*, vol.4, pp.1004–1007, IEEE, 2008.

[66] Q. Jiang, L. Zhang, Y. Zhu, and Y. Zhang, "Larger is better: Seed selection in link-based anti-spamming algorithms," in *Proceedings of the 17th international conference on World Wide Web*, pp.1065–1066, ACM, 2008.

[67] X. Zhang, B. Han, and W. Liang, "Automatic seed set expansion for trust propagation based anti-spamming algorithms," in *Proceedings of the eleventh international workshop on Web information and data management*, pp.31–38, ACM, 2009.

[68] Z. Gyongyi, P. Berkhin, H. Garcia-Molina, and J. Pedersen, "Link spam detection based on mass estimation," in *Proceedings of the 32nd international conference on Very large data bases*, pp.439–450, VLDB Endowment, 2006.

[69] L. Zhao, Q. Jiang, and Y. Zhang, "From good to bad ones: Making spam detection easier," in *Computer and Information Technology Workshops, 2008. CIT Workshops 2008. IEEE 8th International Conference on*, pp.129–134, IEEE, 2008.

[70] Y. Zhang, Q. Jiang, L. Zhang, and Y. Zhu, "Exploiting bidirectional links: making spamming detection easier," in *Proceedings of the 18th ACM conference on Information and knowledge management*, pp.1839–1842, ACM, 2009.

[71] Y. Zhang, Q. Jiang, L. Zhang, and Y. Zhu, "Deeply exploiting link structure: Setting a tougher life for spammers," tech. rep., Technical report, Peking University, 2009. http://www. cis. pku. edu. cn/faculty/system/zhangyan/papers/CPV. pdf, 2009.

[72] X. Zhang, Y. Wang, N. Mou, and W. Liang, "Propagating both trust and distrust with target differentiation for combating web spam.," in *AAAI*, 2011.

[73] X. Liu, Y. Wang, S. Zhu, and H. Lin, "Combating web spam through trust–distrust propagation with confidence," *Pattern Recognition Letters*, vol.34, no.13, pp.1462–1469, 2013.

[74] Y. Liu, B. Gao, T.-Y. Liu, Y. Zhang, Z. Ma, S. He, and H. Li, "Browserank: letting web users vote for page importance," in *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pp.451–458, ACM, 2008.

[75] B. Poblete, C. Castillo, and A. Gionis, "Dr. searcher and mr. browser: a unified hyperlink-click graph," in *Proceedings of the 17th ACM conference on Information and knowledge management*, pp.1123–1132, ACM, 2008.

[76] Y. Liu, M. Zhang, S. Ma, and L. Ru, "User behavior oriented web spam detection," in *Proceedings of the 17th international conference on World Wide Web*, pp.1039–1040, ACM, 2008.

[77] B. Zhou, J. Pei, and Z. Tang, "A spamicity approach to web spam detection.," in *SDM*, pp.277–288, SIAM, 2008.

[78] B. Zhou and J. Pei, "Link spam target detection using page farms," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol.3, no.3, p.13, 2009.

[79] P. Zhou, "Osd: An online web spam detection system," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '09, 2009.

[80] B. Zhou and J. Pei, "Sketching landscapes of page farms.," in *SDM*, pp.593–598, SIAM, 2007.

[81] Y. Yang and J. O. Pedersen, "A comparative study on feature selection in text categorization," in *ICML*, vol.97, pp.412–420, 1997.

[82] S. Webb, J. Caverlee, and C. Pu, "Predicting web spam with http session information," in *Proceedings of the 17th ACM conference on Information and knowledge management*, pp.339–348, ACM, 2008.

[83] K. Chellapilla and A. Maykov, "A taxonomy of javascript redirection spam," in *Proceedings of the 3rd international workshop on Adversarial information retrieval on the web*, pp.81–88, ACM, 2007.

[84] K. M. Svore, Q. Wu, C. J. Burges, and A. Raman, "Improving web spam classification using rank-time features," in *Proceedings of the 3rd international workshop on Adversarial information retrieval on the web*, pp.9–16, ACM, 2007.

[85] F. Radlinski, "Addressing malicious noise in clickthrough data," in *Learning to Rank for Information Retrieval Workshop at SIGIR*, vol.2007, 2007.

[86] Z. Dou, R. Song, X. Yuan, and J.-R. Wen, "Are click-through data adequate for learning web search rankings?," in *Proceedings of the 17th ACM conference on Information and knowledge management*, pp.73–82, ACM, 2008.

[87] R. Bhattacharjee and A. Goel, "Algorithms and incentives for robust ranking," in *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pp.425–433, Society for Industrial and Applied Mathematics, 2007.

[88] G.-G. Geng, C.-H. Wang, Q.-D. Li, L. Xu, and X.-B. Jin, "Boosting the performance of web spam detection with ensemble under-sampling classification," in *Fuzzy Systems and Knowledge Discovery, 2007. FSKD 2007. Fourth International Conference on*, vol.4, pp.583–587, IEEE, 2007.

[89] M. Mahmoudi, A. Yari, and S. Khadivi, "Web spam detection based on discriminative content and link features," in *Telecommunications (IST), 2010 5th International Symposium on*, pp.542–546, IEEE, 2010.

[90] G.-G. Geng, X.-B. Jin, X.-C. Zhang, and D.-X. Zhang, "Evaluating web content quality via multi-scale features," *arXiv preprint arXiv:1304.6181*, 2013.

[91] H. A. Wahsheh, M. N. Al-Kabi, and I. M. Alsmadi, "A link and content hybrid approach for arabic web spam detection," *International Journal of Intelligent Systems and Applications (IJISA)*, vol.5, no.1, pp.30–43, 2013.

[92] H. Liu and R. Setiono, "Chi2: Feature selection and discretization of numeric attributes," in *2012 IEEE 24th International Conference on Tools with Artificial Intelligence*, pp.388–388, IEEE Computer Society, 1995.

[93] M. A. Hall, "Correlation-based feature selection for discrete and numeric class machine learning," in *Proceedings of the Seventeenth International Conference on Machine Learning*, ICML '00, (San Francisco, CA, USA), pp.359–366, Morgan Kaufmann Publishers Inc., 2000.

[94] Z. Jia, W. Li, W. Gao, and Y. Xia, "Research on web spam detection based on support vector machine," in *Communication Systems and Network Technologies (CSNT), 2012 International Conference on*, pp.517–520, IEEE, 2012.

[95] K. L. Goh, A. K. Singh, and K. H. Lim, "Multilayer perceptrons neural network based web spam detection application," in *Signal and Information Processing (ChinaSIP), 2013 IEEE China Summit & International Conference on*, pp.636–640, IEEE, 2013.

[96] M. Erdélyi, A. Garzó, and A. A. Benczúr, "Web spam classification: A few features worth more," in *Proceedings of the 2011 Joint WICOW/AIRWeb Workshop on Web Quality*, WebQuality '11, (New York, NY, USA), pp.27–34, ACM, 2011.

[97] L. Araujo and J. Martinez-Romo, "Web spam detection: new classification features based on qualified link analysis and language models," *Information Forensics and Security, IEEE Transactions on*, vol.5, no.3, pp.581–590, 2010.

[98] J. Abernethy, O. Chapelle, and C. Castillo, "Web spam identification through content and hyperlinks," in *Proceedings of the 4th international workshop on Adversarial information retrieval on the web*, pp.41–44, ACM, 2008.

[99] C. Castillo, D. Donato, L. Becchetti, P. Boldi, S. Leonardi, M. Santini, and S. Vigna, "A reference collection for web spam," in *ACM Sigir Forum*, vol.40, pp.11–24, ACM, 2006.

[100] T. Urvoy, E. Chauveau, P. Filoche, and T. Lavergne, "Tracking web spam with html style similarities," *ACM Trans. Web*, vol.2, pp.3:1–3:28, Mar. 2008.

[101] "The gzip home page," `http://www.gzip.org`, accessed September, 2013.

[102] S. K. M. Wong, W. Ziarko, V. V. Raghavan, and P. Wong, "On modeling of information retrieval concepts in vector spaces," *ACM Transactions on Database Systems (TODS)*, vol.12, no.2, pp.299–321, 1987.

[103] Z. S. Harris, "Distributional structure.," *Word*, 1954.

[104] R. A. Horn and C. R. Johnson. *Matrix analysis*. Cambridge university press, 2012.

[105] "Web spam challenge ii: Small corpus," `http://webspam.lip6.fr/wiki/pmwiki.php?n=Main.PhaseIICorpora/`, Provided by Ludovic Denoyer etc, University of Paris 6, France, 2006.

[106] "Yahoo! research: Web spam collections," `http://barcelona.research.yahoo.net/webspam/datasets/`, Crawled by the Laboratory of Web Algorithmics, University of Milan, http://law.dsi.unimi.it, 2007.

[107] "The lemur project," http://www.lemurproject.org, accessed September 2013.

[108] "Weka 3: Data mining software in java," http:///www.cs.waikato.ac.nz/ml/weka, accessed November 2013.

[109] "Libsvm – a library for support vector machines," http:///www.csie.ntu.edu.tw/~cjlin/libsvm, accessed November 2013.

[110] V. Nikulin, "Web-mining with wilcoxon-based feature selection, ensembling and multiple binary classifiers," in *Proceedings of the ECML/PKDD*, 2010.

[111] F. J. Ortega, J. A. Troyano, F. L. Cruz, and C. G. Vallejo, "Polarityspam: Propagating content-based information through a web-graph to detect web-spam," *International Journal of Innovative Computing, Information and Control*, vol.8, no.5, 2012.

واژه‌نامه انگلیسی به فارسی

متن پیوند ..... anchor text
پس‌رو ..... backward
حذف پس‌رو ..... backward elimination
کیف کلمات هرز ..... bag-of-spam-words
کیف کلمات ..... bag-of-words
پایه ..... baseline
سوگیری ..... bias
دوجهته ..... bidirectional
دودویی ..... binary
خط مرزی ..... border line
رده‌بندی ..... classification
رده‌بند ..... classifier
تلیک ..... click
هرز تلیک ..... click spam
در سمت سرویس‌گیرنده ..... client-side
خوشه‌بندی ..... clustering
ضریب اعتماد ..... confidence factor
همگرایی ..... convergence
شباهت کسینوسی ..... cosine similarity
عامل میرایی ..... damping factor
مجموعه داده‌ای ..... dataset
درخت تصمیم ..... decision tree
شناسایی ..... detection
انجمن‌های گفتگو ..... discussion forums
عدم اعتبار ..... distrust
گوناگونی ..... diversity
دامنه ..... domain
پویا ..... dynamic
یال ..... edge
رایانامه ..... email
منفی‌ـ کاذب ..... false negative
مثبت‌ـ کاذب ..... false positive
ویژگی ..... feature
انتخاب ویژگی ..... feature selection
بازخورد ..... feedback
انگشت‌نگاری ..... fingerprinting
انجمن ..... forum
پیش‌رو ..... forward
حریصانه ..... greedy
نام میزبان ..... host name
در سطح میزبان ..... host-level
نمایه‌سازی ..... indexing
بازیابی اطلاعات ..... Information Retrieval (IR)
معکوس ..... inverse
ضریب جاکارد ..... Jaccard coeficient
جاوا اسکریپت ..... javaScript
انباشتگی کلیدواژه‌ها ..... keyword stuffing
کلیدواژه‌ها ..... keywords
اعتبار سنجی متقاطع kـ بخشی ..... k-fold cross validation
برچسب ..... label
مدل زبانی ..... language model
پیوند ..... link
دهکده پیوندی ..... link farm
یادگیری ماشین ..... machine learning

بدافزار ..... malware
ابربرچسب ..... Meta tag
پردازش زبان طبیعی . Natural Language Processing (NLP)
گره ..... node
غیرنشانه‌گذاری ..... non-markup
هنجارسازی ..... normalization
به‌هنجار ..... normalized
برخط ..... online
برون‌هشته ..... outlier
دهکده صفحه ..... page farm
در سطح صفحه ..... page-level
شخصی‌سازی شده ..... personalized
بسترهای نرم‌افزاری ..... platforms
دقت ..... precision
اصلی ..... primitive
پرس‌وجو ..... query
جنگل تصادفی ..... random forest
رتبه ..... rank
رتبه‌دهی ..... ranking
بی‌درنگ ..... real-time
فراخوانی ..... recall
بهینه‌سازی موتور جست‌وجو ..... Search Engine Optimization (SEO)
موتورهای جست‌وجو ..... search engines
هسته ..... seed
نیمه‌سرپرست ..... semi-supervised
در سمت سرویس‌دهنده ..... server-side
نشست ..... session
هرز ..... spam
ضریب هرز ..... spam factor
هرزنویسان ..... spammers
ایستا ..... static
مدل‌های زبانی آماری ... statistical language models
ایست‌واژه ..... stopword
قویا متصل ..... strongly connected
باسرپرست ..... supervised
بردارهای پشتیبان ..... support vectors
سامانه ..... system
مدل موضوعی ..... topic model
موضوعی ..... topical
داده‌های آموزش ..... train data
منفی‌ـ‌صحیح ..... true negative
مثبت‌ـ‌صحیح ..... true positive
اعتبار ..... trust
بدون‌سرپرست ..... unsupervised
فضای برداری ..... vector space
هرز وب ..... web spam
هرزنویسی وب ..... web spamming
وب‌نوشت ..... weblog
وب‌گاه ..... website

واژه‌نامه فارسی به انگلیسی

ابربرچسب ..... Meta tag
اصلی ..... primitive
اعتبار ..... trust
اعتبار سنجی متقاطع kـ بخشی ..... k-fold cross validation
انباشتگی کلیدواژه‌ها ..... keyword stuffing
انتخاب ویژگی ..... feature selection
انجمن ..... forum
انجمن‌های گفتگو ..... discussion forums
انگشت‌نگاری ..... fingerprinting
ایستا ..... static
ایست‌واژه ..... stopword
بازخورد ..... feedback
بازیابی اطلاعات ..... Information Retrieval (IR)
باسرپرست ..... supervised
بدافزار ..... malware
بدون‌سرپرست ..... unsupervised
برچسب ..... label
برخط ..... online
بردارهای پشتیبان ..... support vectors
برون‌هشته ..... outlier
بسترهای نرم‌افزاری ..... platforms
به‌هنجار ..... normalized
بهینه‌سازی موتور جست‌وجو ..... Search Engine Optimization (SEO)
بی‌درنگ ..... real-time
پایه ..... baseline
پردازش زبان طبیعی ..... Natural Language Processing (NLP)
پرس‌وجو ..... query
پس‌رو ..... backward
پویا ..... dynamic
پیش‌رو ..... forward
پیوند ..... link
تلیک ..... click
جاوا اسکریپت ..... javaScript
جنگل تصادفی ..... random forest
حذف پس‌رو ..... backward elimination
حریصانه ..... greedy
خط مرزی ..... border line
خوشه‌بندی ..... clustering
داده‌های آموزش ..... train data
دامنه ..... domain
در سطح صفحه ..... page-level
در سطح میزبان ..... host-level
در سمت سرویس‌دهنده ..... server-side
در سمت سرویس‌گیرنده ..... client-side
درخت تصمیم ..... decision tree
دقت ..... precision
دوجهته ..... bidirectional
دودویی ..... binary
دهکده پیوندی ..... link farm
دهکده صفحه ..... page farm
رایانامه ..... email
رتبه ..... rank

رتبه‌دهی ranking
رده‌بند classifier
رده‌بندی classification
سامانه system
سوگیری bias
شباهت کسینوسی cosine similarity
شخصی‌سازی شده personalized
شناسایی detection
ضریب اعتماد confidence factor
ضریب جاکارد Jaccard coeficient
ضریب هرز spam factor
عامل میرایی damping factor
عدم اعتبار distrust
غیرنشانه‌گذاری non-markup
فراخوانی recall
فضای برداری vector space
قویا متصل strongly connected
کلیدواژه‌ها keywords
کیف کلمات bag-of-words
کیف کلمات هرز bag-of-spam-words
گره node
گوناگونی diversity
متن پیوند anchor text
مثبت‌ـ صحیح true positive
مثبت‌ـ کاذب false positive
مجموعه داده‌ای dataset
مدل زبانی language model
مدل موضوعی topic model
مدل‌های زبانی آماری statistical language models
معکوس inverse
منفی‌ـ صحیح true negative
منفی‌ـ کاذب false negative
موتورهای جست‌وجو search engines
موضوعی topical
نام میزبان host name
نشست session
نمایه‌سازی indexing
نیمه‌سرپرست semi-supervised
وب‌گاه website
وب‌نوشت weblog
ویژگی feature
هرز spam
هرز تلیک click spam
هرز وب web spam
هرزنویسان spammers
هرزنویسی وب web spamming
هسته seed
همگرایی convergence
هنجارسازی normalization
یادگیری ماشین machine learning
یال edge

Abstract

In recent years, due to the increasing amount of data available on the internet, the use of search engines to retrieve relevant information from the World Wide Web has become pervasive. Among the huge number of websites, the ones which succeed to appear more frequently and in higher ranks of search engine results would receive more visitors. So, spammers struggle to achieve a higher than deserved rank for their websites using some illegal techniques called web spamming. Although various methods have been used for combatting web spamming, we could basically categorize them into three groups: content-based methods, link-based methods, and the methods based on miscellaneous data. In this thesis, we focus on content-based and link-based methods, and also their combination.

Despite the existence of many spam detection methods, the search engines do not perform well in detecting Persian spam websites. Thus, in this thesis, after preparing a corpus of spam and non-spam Persian websites, we analyze the effectiveness of many previously proposed content-based features on detecting Persian spam websites. To improve the performance of classification, we present a number of new content-based features and examine a number of feature selection method. As another approach, we propose a new Persian spam detection system which uses an improved version of bag-of-words model and has better performance in detecting Persian web spam. Due to the prevalence of link-based spamming methods, we analyze some of these methods and propose two new algorithms which do not have the weaknesses of previous methods. In the first algorithm, to improve the process of label propagation, we use three mechanisms: optimized seed selection, edge weighting, and seed expansion. In the second algorithm, we improve the quality of websites ranking, using label propagation in both forward and backward directions. Finally, we propose a combined method, which uses the content-based probability of being spam (non-spam) to propagate the spam (non-spam) score of websites. Using this method, we increase the performance of ranking websites.

Finally, to evaluate the proposed methods and compare their performance with the existing methods for this task, we have conducted several experiments on different datasets. Experiment results indicate that the proposed methods have a good performance in detecting web spam.

Keywords: *Spamming, Web Spam, Spam Detection, Label Propagation, Content-Based Features*

University of Tehran
School of Electrical and Computer Engineering

Detecting Persian Spam Web Pages

By
Elahe Rabbani

Supervisors:
Dr. Azadeh Shakery

A thesis submitted to the Graduate Studies Office
in partial fulfillment of the requirements
for the degree of Master of Science

in

Computer Engineering

September 2014