مقاله پژوهشی

تشخیص صفحات اسپم با استفاده از الگوریتم XGBoost

ریحانه رشیدپور و علی‌محمد زارع بیدکی

چکیده: امروزه موتورهای جستجو دروازه ورود به وب هستند. با افزایش محبوبیت وب، تلاش برای بهره‌برداری تجاری، اجتماعی و سیاسی از وب نیز افزایش یافته و در نتیجه تشخیص یک محتوای خوب از اسپم برای موتورهای جستجو دشوار شده است. مفهوم اسپم وب نخستین بار در سال ۱۹۹۶ معرفی شد و خیلی زود به عنوان یکی از چالش‌های کلیدی برای صنعت موتور جستجو شناخته شد. پدیده اسپم اساساً به این دلیل اتفاق می‌افتد که بخش قابل توجهی از مراجعات به صفحه وب از موتور جستجو می‌آیند و کاربران تمایل به بررسی اولین نتایج جستجو دارند. هدف از شناسایی صفحات اسپم این است که این صفحات با استفاده از استراتژی‌های فریب قادر به کسب رتبه بالا نباشند. تلاش ما ارائه روشی مؤثر در شناسایی صفحات اسپم و در نتیجه کاهش حضور اسپم در نتایج اول جستجوست. در این مقاله دو روش برای مقابله با اسپم وب پیشنهاد شده است. روش اول به نام XGspam صفحات اسپم را بر اساس الگوریتم یادگیری XGBoost با دقت ۹۴٫۲۷٪ شناسایی می‌کند. در روش دوم به نام XGSspam راهکاری برای چالش نامتوازن‌بودن داده‌های وب با استفاده از ترکیب الگوریتم بیش‌نمونه‌برداری SMOTE با مدل دسته‌بندی XGBoost ارائه شده که به دقت ۹۵٫۴۴٪ در شناسایی صفحات اسپم می‌رسد.

کلیدواژه: **اسپم وب، الگوریتم دسته‌بندی XGBoost، متوازن‌سازی داده، یادگیری ماشین.**

۱- مقدمه

اولین نسخه وب در سال ۱۹۸۹ توسط برنرز لی به وجود آمد [۱]. با گسترش وب، کاربردهای متنوعی در زمینه‌های مختلف برای وب ایجاد شد و روزبه‌روز بر تعداد کاربران وب افزوده شد. با گذشت زمان، استفاده بهینه از حجم عظیم اطلاعات وب دشوار شد. امروزه حدود ۵٫۸۳ بیلیون[1] صفحه وب وجود دارد [۲]. موتورهای جستجو با جمع‌آوری و پردازش محتوای ارائه‌شده در وب، امکان دستیابی راحت و سریع به محتوای مورد نظر را برای کاربران فراهم آوردند. Lycos، نخستین موتور جستجوی تجاری در ۱۹۹۵ وارد دنیای وب شد. از همان زمان مبارزه با صفحات اسپم تبدیل به یکی از مباحث مورد علاقه مجامع دانشگاهی شد [۳].

موتور جستجو صفحات وب را بر مبنای کیفیت آنها و میزان ارتباطشان با پرس‌وجوی کاربر رتبه‌بندی می‌کند؛ به‌طوری که معمولاً اولین نتایج حاصل از جستجو بهترین و مرتبط‌ترین نتایج هستند. اما متأسفانه وجود عوامل مزاحمی به نام اسپم وب مانع دقت کافی نتایج جستجو و در نتیجه مانع رضایت کامل کاربران از موتور جستجو می‌شوند. افراد سودجو با هدف بهره‌برداری تجاری از وب، اقدام به ایجاد وب‌سایت‌های اسپم می‌کنند و در ساخت این صفحات از تکنیک‌های متفاوتی استفاده می‌کنند. حضور اسپم در اولین نتایج جستجو منجر به اتلاف وقت کاربران و نارضایتی آنان خواهد شد. همچنین خزش، نمایه‌سازی و رتبه‌بندی صفحات اسپم، منجر به تحمیل هزینه محاسباتی زیادی به شرکت‌های موتور جستجو خواهد شد. به همین دلیل مقابله با اسپم وب یکی از چالش‌های اصلی موتورهای جستجو است.

اسپمرها از تکنیک‌های متنوع مبتنی بر محتوا و پیوند برای افزایش رتبه صفحاتشان استفاده می‌کنند. در مقابل موتورهای جستجو با شناسایی تکنیک‌های اسپم‌ساز، الگوریتم‌های رتبه‌بندی جدید ارائه می‌دهند. متقابلاً اسپمرها برای فرار از شناسایی و حذف صفحاتشان از تکنیک‌های دیگری استفاده خواهند کرد. با توجه به تنوع صفحات اسپم، راهکارهای متفاوتی برای مقابله با آنها ارائه شده است؛ اما به دلیل تنوع روش‌های تقلب اسپم و پویایی اسپمرها در فریب الگوریتم‌های رتبه‌بندی موتور جستجو، تاکنون هیچ کدام از روش‌های مقابله با اسپم کاملاً موفق نبوده‌اند.

در این مقاله با بررسی مجموعه داده WEBSPAM-UK۲۰۰۷ [۴] و با استفاده از تعدادی از ویژگی‌های محاسبه‌شده در مجموعه داده برای صفحات وب به ارائه روشی برای شناسایی صفحات اسپم پرداخته‌ایم. در مجموعه داده مورد استفاده تعداد صفحات اسپم بسیار کمتر از صفحات خوب است؛ تنها حدود ۵٪ از صفحات مجموعه داده اسپم هستند. ما ابتدا به اجرای تعدادی از پرکاربردترین روش‌های ML برای دسته‌بندی صفحات پرداختیم که بهترین دقت دسته‌بندی و کمترین زمان محاسباتی متعلق به الگوریتم XGBoost [۵] بود. سپس با استفاده از تکنیک SMOTE داده‌ها متوازن‌سازی شدند [۶] و مجدداً روش XGBoost بر روی داده‌های متوازن اجرا شد که منجر به بهبود دقت طبقه‌بندی تا ۹۵٪٫۴۴ شد.

در ادامه این مقاله ابتدا در بخش ۲ مفاهیم اولیه مرتبط با این موضوع تعریف گردیده و سپس در بخش ۳ تعدادی از روش‌های پیشین مقابله با اسپم وب مرور شده است. در بخش ۴ روش پیشنهادی تشریح شده و نهایتاً در بخش ۵ نتایج حاصل بیان شده‌اند.

۲- تعاریف اولیه

در این بخش، مفاهیم و تعاریف اولیه استفاده‌شده در این مقاله به‌طور خلاصه معرفی می‌شوند.

۲-۱ گراف وب

وب به عنوان یک گراف $G(V,E)$ مدل‌سازی می‌شود [۷] که در آن گره‌ها (V) نشان‌دهنده صفحات وب و یال‌های وزن‌دار جهت‌دار نشان‌دهنده پیوندهای بین صفحات هستند. اگر صفحه p_i چندین پیوند به صفحه p_j داشته باشد، همه در یک پیوند مجتمع می‌شوند $(i,j) \in E$. حلقه روی یک صفحه مجاز نیست. تعداد صفحاتی که از

این مقاله در تاریخ ۲۹ خرداد ماه ۱۴۰۳ دریافت و در تاریخ ۱۰ شهریور ماه ۱۴۰۳ بازنگری شد.

ریحانه رشیدپور (نویسنده مسئول)، دانشکده مهندسی کامپیوتر، دانشگاه یزد، یزد، ایران، (email: rashidpour@stu.yazd.ac.ir).

علی‌محمد زارع بیدکی، دانشکده مهندسی کامپیوتر، دانشگاه یزد، یزد، ایران، (email: alizareh@yazd.ac.ir).

1. 1 Billion = 10^{12}

صفحه p_i پیوند دریافت می‌کنند درجه خروجی و تعداد صفحاتی که به p_i پیوند دارند، درجه ورودی آن صفحه می‌باشد. هر یال $(i,j) \in E$ می‌تواند یک وزن w_{ij} داشته باشد. یک استراتژی رایج برای اختصاص وزن به یال‌ها به صورت (۱) است که در آن $out(p_i)$ درجه خروجی صفحه i می‌باشد

$$W_{ij} = \frac{1}{out(p_i)} \quad (1)$$

۲-۲ اسپم

بخش بزرگی از بازدیدهای یک وب‌سایت از موتورهای جستجو سرچشمه می‌گیرد و بیشتر کاربران روی چند نتیجه اول جستجو کلیک می‌کنند؛ بنابراین برای دستکاری نتایج جستجو با ایجاد صفحات اسپم، انگیزه اقتصادی وجود دارد. صفحاتی که مستقل از شایستگی واقعیشان سعی بر کسب امتیاز بالایی دارند، اسپم هستند. به عبارت دیگر هر تلاشی برای فریب موتور جستجو اسپم است [۸].

۲-۳ انواع اسپم

اسپم وب را می‌توان به سه دسته کلی تقسیم کرد. اسپم محتوایی که در آن محتوای صفحه وب برای رسیدن به رتبه بالاتر دستکاری می‌شود. برای تشخیص این نوع اسپم از روش‌های مبتنی بر محتوای صفحه استفاده می‌شود که تعیین می‌کند آیا یک صفحه وب بر اساس مشخصات محتوایی صفحه، اسپم است یا نه. اسپم پیوند که در آن پیوندهای بین صفحات و در نتیجه گراف وب دستکاری می‌شود. برای تشخیص این نوع اسپم از روش‌های مبتنی بر پیوند استفاده می‌شود که اسپم وب را بر اساس ارتباطات بین صفحات وب تشخیص می‌دهد. اسپم رفتاری که عمدتاً مربوط به کلیک‌های برنامه‌ریزی‌شده برای اهداف سودجویانه است و روش‌های تشخیص این نوع اسپم بر اساس داده‌های رفتار کاربر، کلیک‌ها و غیره انجام می‌شود.

۳- روش‌های مقابله با اسپم وب

تاکنون تلاش‌های بسیاری در زمینه شناسایی و مقابله با صفحات اسپم صورت گرفته و در ادامه، نمونه‌هایی از الگوریتم‌های متنوع مقابله با اسپم وب آمده است.

دولاس و همکاران در [۹] تعدادی ویژگی محتوایی جدید معرفی نموده‌اند و پس از آن با به‌کارگیری روش‌های یادگیری ماشین، این ویژگی‌ها را برای ایجاد یک الگوریتم تشخیص اسپم کارآمد ترکیب کرده‌اند. این پژوهش بر روی زیرمجموعه‌ای از صفحات خزش‌شده توسط MSN Search انجام شده است. یک نمونه تصادفی شامل ۱۷۱۶۸ صفحه به طور دستی بررسی و برچسب‌گذاری شده است. حدود ۱۳/۸٪ صفحات برچسب اسپم و ۸۶/۲٪ صفحات برچسب نرمال خوردند. نهایتاً همه این ویژگی‌ها در یک مدل کلاس‌بندی در چارچوب‌های C۴.۵، وزن‌دهی[1] و ترکیب[2] شدند. دقت حاصل در شناسایی درست صفحات اسپم ۸۶٪/۲ حاصل شد.

بکتی و همکاران در [۱۰] توابع استهلاک عمومی برای تشخیص اسپم وب را بررسی کردند و الگوریتم Truncated PageRank را پیشنهاد دادند. ایده اصلی این الگوریتم آن است که صفحات اسپم در فاصله‌های کوتاه تعداد زیادی پیونددهنده دارند؛ در حالی که این تعداد در فاصله‌های طولانی‌تر کمتر از حد مورد انتظار است. در نتیجه از تابع استهلاکی استفاده شد که مشارکت مستقیم از چند سطح اول پیوند ورودی را نادیده بگیرد. مجموعه داده مورد استفاده، مجموعه‌ای از ۱۸/۵ میلیون صفحه از دامنه uk. است که در سال ۲۰۰۲ دانلود شد. صفحات در ۹۸۴۵۲ میزبان مختلف قرار داشتند. با توجه به حجم زیاد این مجموعه، هاست‌ها را به جای صفحات جداگانه دسته‌بندی کردند. نویسندگان نمونه‌ای از ۵۷۵۰ میزبان (۵/۹٪) را به صورت دستی بررسی کردند و به‌طور دستی برچسب زدند که ۳۱/۷٪ از صفحات وب مجموعه را پوشش می‌داد. دقت تشخیص اسپم وب در این روش ۸۰ درصد می‌باشد که معادل با دقت بهترین دسته‌بندی‌های اسپم محتوایی است.

بایزا یاتیس و همکاران در [۱۱] با الهام از نمایش PageRank مفهوم رتبه تابعی[3] را پیشنهاد داده‌اند که یک تعمیم از PageRank با توابع استهلاک[4] متنوع می‌باشد. آنها رتبه‌بندی را در یک فرمول کلی (۲) سنجیده‌اند که در آن P ماتریس نرمال لینک‌ها و N تعداد نودهاست. $Damping(j)$ تابع استهلاک برای نود j و M^T ترانهاده ماتریس همسایگی گراف است

$$P = \frac{1}{N}\sum_{j=0}^{\infty} damping(j)(M^T)^j \quad (2)$$

و این نظریه را اثبات کردند که هر تابع استهلاک (مثلاً تابعی که در آن مجموع استهلاک‌ها ۱ می‌شود) بیانگر یک رتبه‌بندی کارکردی نرمال خوش‌ساخت است. آنها توابع استهلاک توانی (PageRank)، خطی و غیره را مطالعه کردند و روش مؤثری برای محاسبه رتبه پیشنهاد دادند.

در [۱۲]، یک روش جدید مبتنی بر تمایز صفحات وب (DPR) به منظور بهبود مضرات الگوریتم PageRank کلاسیک در تخصیص وزن پیوندها به طور مساوی و نادیده‌گرفتن اعتبار صفحات وب پیشنهاد شده است. همچنین در [۱۳]، یک الگوریتم ضد اعتماد ناهمزمان توسعه داده شده است تا به طور قابل توجهی تعداد عملیات حسابی را در مقایسه با الگوریتم سنتی TrustRank [۱۴] بدون کاهش عملکرد در تشخیص هرزنامه وب کاهش دهد. الگوریتم BadRank نیز ایده محاسبه میزان بدی یک صفحه با استفاده از محاسبه PageRank معکوس را پیشنهاد می‌دهد [۱۵].

لیو و همکاران روشی را برای انتخاب ویژگی‌ها بر اساس الگوریتم جنگل تصادفی پیشنهاد دادند. این روش بر اساس مجموعه داده WEBSPAM-UK۲۰۰۷ طراحی شده است. در ابتدا با در نظر گرفتن ویژگی‌های مرتبط با صفحه اصلی هاست، با پسوند hp. تعداد ویژگی‌ها به ۱۰۶ ویژگی کاهش یافته است. سپس از روش حذف معکوس برای انتخاب از میان این ۱۰۶ ویژگی استفاده شده است. این روش تأثیر بر نتیجه دسته‌بندی را پس از حذف یک مجموعه از ویژگی‌ها اندازه‌گیری می‌کند و نه حذف یک ویژگی به تنهایی. در این پژوهش ۷ ویژگی جدید استخراج شده و بر اساس روش انتخاب ویژگی ارائه‌شده در [۸]، تعداد ۱۴ ویژگی از ویژگی‌های موجود در مجموعه داده انتخاب شده است. در مجموع با استفاده از ۲۱ ویژگی به بهبود امتیاز F۱ تا ۹۲٪ دست یافته است [۷].

در [۱۶] روشی برای شناسایی صفحات اسپم از دیدگاه کاربر با بررسی

1. Boosting
2. Bagging
3. Functional Rank
4. Damping

صفحاتی که نسخه خزشگر آنها با نسخه نمایش‌داده‌شده به کاربر متفاوت است، مثل اسپم محتوای پنهان و اسپم تغییر مسیر معرفی شده است. در این روش از صفحه وب نمایش‌داده‌شده به کاربر، عکس گرفته شده و سپس از CNN برای تجزیه و تحلیل و طبقه‌بندی تصاویر استفاده شده است. نویسندگان ابتدا یک مجموعه داده را برای ارزیابی روش پیشنهادی جمع‌آوری کردند و نشان دادند که این روش با دقت ۹۱٪ نسبت به روش‌های یادگیری ماشین سنتی دقت بهتری دارد. سپس به مدت ۳ ماه الگوریتم پیشنهادی را بر روی صفحات وب واقعی اجرا کردند. سرورهای نصب‌شده در ۵ ماشین از ۲۰ میلیون صفحه وب اسکرین‌شات تهیه کردند و سپس به بررسی این صفحات پرداختند. این روش شناسایی صفحات اسپم، تمام تکنیک‌های اسپمرها در لایه‌های میانی را خنثی می‌کند. هرچند آموزش این روش زمان‌بر است، اما آزمایش آن کارامد است و به نتایج بسیار مطلوبی دست یافته است.

ژوانگ و همکاران در [۱۷] یک روش جدید یادگیری مبتنی بر اولویت[1] را با هدف کاهش رتبه[2] صفحات اسپم در رتبه‌بندی مبتنی بر انتشار امتیاز پیشنهاد داده‌اند. این مقاله یکی از اولین مطالعاتی است که از روش یادگیری برای رتبه‌بندی[3] برای مقابله با مشکل اسپم وب استفاده کرده است. نویسندگان به بررسی روش‌های رتبه‌بندی صفحات وب از طریق انتشار امتیاز در گراف وب پرداخته‌اند. بدیهی است الگوریتم‌هایی که بتوانند با کاهش رتبه صفحات اسپم از حضور آنها در نتایج اول جستجو جلوگیری کنند، موفق‌تر خواهند بود. ایشان برای ارزیابی میزان موفقیت الگوریتم‌های رتبه‌بندی، یک متریک جدید به نام امتیاز کاهش رتبه[4] ارائه دادند و به بررسی و مقایسه روش پیشنهادی با سایر الگوریتم‌های رتبه‌بندی بر اساس متریک پیشنهادی پرداختند. در مقایسه با الگوریتم‌های کاهش رتبه صفحات اسپم[5]، مبتنی بر انتشار امتیاز ۰/۷٪ بهبود نسبی در مجموعه داده WEBSPAM-UK۲۰۰۶ و ۷/۳٪ بهبود نسبی در مجموعه داده WEBSPAM-UK۲۰۰۷ از نظر امتیاز کاهش رتبه صفحات اسپم به دست آمده است.

رفتار کاربران یکی از پارامترهای مهم مورد توجه اسپمرها برای طراحی تکنیک‌های اسپم‌ساز است. یک ارتباط جالب بین کلمات کلیدی موجود در پرس‌وجوی کاربر و URLهای سایت‌های اسپم وجود دارد. در [۱۸] این نکته مورد توجه قرار گرفت و در نتیجه یک چارچوب تشخیص اسپم با هدف تحلیل داده مربوط به کلیک در سطح سایت و صفحه، پیشنهاد شد و الگوریتم انتشار برچسب نامیده شد که در عمل با توجه به پیمایش سراسری گراف وب خزش‌شده، نسبت به الگوریتم‌های PageRank و TrustRank نتایج بهتری داشته است.

اسپمرها حتی به سایت‌های شبکه اجتماعی هم صدمه زده‌اند و تکنیک‌های تشخیص اسپم مجزا برای شبکه‌های اجتماعی پیشنهاد شده‌اند. ایده مرور محصولات توسط کاربران، منجر به اسپم مرور شد که برای شناسایی آن تکنیک‌هایی ارائه شده است [۱۹]. مطالعات بسیاری برای شناسایی اسپم ایمیل، اسپم مرور و ... انجام شده که در چارچوب این مقاله نمی‌گنجد.

1. Preference Based
2. Demotion
3. Learning to Rank (LTR)
4. Demotion Score
5. Web Spam Demotion Algorithms

٤- دسته‌بندی صفحات وب

هدف این مقاله، یافتن راه‌حلی برای مسئله اسپم وب است. صفحات اسپم ویژگی‌هایی دارند که آنها را از سایر صفحات وب متمایز می‌کند؛ مثلاً می‌توان به تکرار کلمات و کاراکترهای خاص در متن صفحه یا در URL صفحات اسپم و یا وجود لینک‌های گروهی به سایر صفحات اسپم اشاره کرد. در این پژوهش تلاش می‌شود که با تشخیص صفحات اسپم و حذف یا جریمه آنها، حضور این نوع صفحات در نتایج اول جستجوی وب به حداقل برسد. اطلاعاتی که می‌توان از هر صفحه وب کسب کرد، شامل اجزای درون صفحه وب مثل متن و تگ‌های HTML است که ویژگی‌های مبتنی بر محتوا نامیده می‌شوند و نیز ویژگی‌های مرتبط با ارتباطات یک صفحه با سایر صفحات وب که ویژگی‌های مبتنی بر پیوند نامیده می‌شوند. بر اساس این دو گروه اصلی ویژگی‌های صفحه و ساختار گراف وب می‌توان پردازش‌هایی انجام داد و ویژگی‌های پیچیده‌تری را برای صفحه محاسبه کرد؛ برای نمونه امتیازات PageRank [۲۰]، TrustRank [۱۴] و TruncatedPageRank [۱۰] که در ابتدا به عنوان الگوریتم‌های رتبه‌بندی وب معرفی شدند، در حال حاضر امتیازات حاصل از آنها به عنوان یک ویژگی صفحه وب مورد استفاده قرار می‌گیرد.

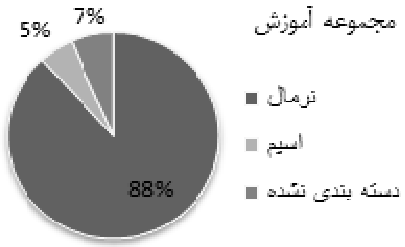
الگوریتم رتبه‌بندی PageRank بر اساس انتشار امتیاز در گراف وب به هر گره (صفحه وب) بر اساس میزان محبوبیت و اعتبار همسایگان آن امتیازی را اختصاص می‌دهد. هرچقدر همسایگان یک صفحه معتبرتر باشند، آن صفحه امتیاز PageRank بیشتری را کسب خواهد کرد و در رتبه‌بندی جایگاه بهتری را خواهد داشت.

الگوریتم TrustRank برای شناسایی صفحات باکیفیت و قراردادن آنها در رتبه‌های بالا، با فرض «یک صفحه خوب با صفحات خوب پیوند دارد»، از یک مجموعه اولیه از صفحات خوب استفاده می‌کند و سپس با دنبال‌کردن لینک‌های خروجی این صفحات به انتشار امتیاز می‌پردازد.

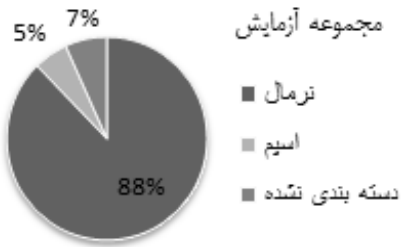
الگوریتم رتبه‌بندی TruncatedPageRank بر مبنای الگوریتم PageRank و با هدف شناسایی و حذف اسپم مزرعه پیوند، ایجاد شده است. مزرعه پیوند نوعی اسپم مبتنی بر پیوند است که در آن صفحات اسپم به تعداد زیاد بین خودشان لینک ایجاد می‌کنند و سعی بر بالابردن امتیاز صفحه‌ای خاص دارند. الگوریتم TruncatedPageRank با حذف پیوند همسایگان بسیار نزدیک یک صفحه (همسایگان با طول مسیر کمتر از n) تا حد زیادی موفق به مقابله با این نوع اسپم شد. به این ترتیب می‌توان ادعا کرد در زمان استفاده از ویژگی‌های مبتنی بر پیوند، ساختار گراف وب و انتشار امتیاز نیز به طور ضمنی در محاسبات دیده می‌شود.

تاکنون پژوهش‌های بسیاری در زمینه شناسایی صفحات اسپم انجام شده است. روش‌های متنوع با تحلیل ویژگی‌های مبتنی بر محتوا و پیوند صفحات وب به دسته‌بندی این صفحات و یا انتشار امتیاز در گراف وب و جریمه صفحات اسپم پرداخته‌اند؛ لیکن به دلیل وجود چالش‌های ماهیتی وب، نظیر پراکنده‌بودن داده‌های گراف وب و نامتوازن‌بودن تعداد صفحات اسپم و نرمال و نظایر آن به دقت مطلوب نرسیده‌اند. این مشکل همچنان مطرح است و هنوز راه‌حل قطعی و کاملی برای آن پیشنهاد نشده است. دسته‌بندی صفحات وب یک راه‌حل مناسب برای رهایی از خسارات ناشی از اسپم وب است. در پژوهش‌های پیشین نیز الگوریتم‌های دسته‌بندی بسیاری برای این منظور به کار گرفته شده‌اند.

در این مقاله تلاش شده تا با یافتن راهکاری برای عبور از موانع موجود، دسته‌بندی دقیق‌تری با استفاده از تعداد کمتری از ویژگی‌ها با سرعت بالاتر و هزینه محاسباتی کمتر انجام شود. در ادامه صفحات وب به دو روش دسته‌بندی شده‌اند:



شکل ۱: توزیع آماری صفحات مجموعه آموزش.



شکل ۲: توزیع آماری صفحات مجموعه آزمایش.

- روش اول که XGspam نامیده می‌شود، شامل استفاده از یک الگوریتم دسته‌بندی سریع و کارا با هدف شناسایی صفحات اسپم است. در این مرحله، مدل دسته‌بندی روی مجموعه داده‌ای از صفحات وب شامل ویژگی‌های محاسبه‌شده و برچسب‌های کلاس، آموزش می‌بیند و سپس روی یک مجموعه بزرگ‌تر، آزمایش و دقت دسته‌بندی محاسبه می‌شود.
- در روش دوم با نام XGSspam برای حل مشکل عدم توازن داده‌ها راهکاری پیشنهاد گردیده و مجدداً مجموعه داده متوازن به دو کلاس اسپم و نرمال دسته‌بندی شده است. ترکیب مدل دسته‌بندی درختی با تکنیک متوازن‌سازی داده‌ها منجر به دقت مطلوبی شده است.

۴-۱ دسته‌بندی XGspam

با توجه به خسارات ناشی از وجود اسپم به شرکت‌های موتور جستجو و به کاربران وب، هدف ما شناسایی صفحات اسپم است. اندازه بسیار بزرگ و روبه‌رشد وب، مستلزم استفاده از یک روش مقیاس‌پذیر، سریع و با هزینه محاسباتی کم است.

با توجه به ویژگی‌های ممتاز روش XGBoost و درخشش آن در رقابت‌های جهانی، به نظر می‌رسد که این روش در حیطه تشخیص صفحات اسپم نیز کارامد باشد. در این مقاله الگوریتم XGBoost برای نخستین بار با هدف دسته‌بندی صفحات وب به کار رفته است.

۴-۱-۱ معرفی الگوریتم XGBoost

XGBoost[1] یک الگوریتم یادگیری ماشین است که از مدل‌های درختی گروهی استفاده می‌کند و با الهام از الگوریتم‌های تقویت درختی (متشکل از درختان تصمیم) از ترکیب چندین درخت رگرسیون[2] به عنوان مدل اصلی خود استفاده می‌کند. در این روش هر برگ درخت (در مثال ما هر صفحه وب) شامل یک امتیاز پیوسته است. برای دسته‌بندی این برگ‌ها از قوانین تصمیم در درخت‌ها استفاده می‌شود. کلاس نهایی هر برگ با تجمیع امتیازات برگ‌های متناظر پیش‌بینی می‌گردد و تلاش می‌شود تا بهترین پیش‌بینی را برای مسائل یادگیری ماشین ارائه کند.

مدل یادگیری ماشین XGBoost یکی از مدل‌های پرطرفدار در رقابت‌های Kaggle و KDD-cup است که تاکنون جوایز بسیاری را به خود اختصاص داده است [۲۱] و [۲۲]. مهم‌ترین فاکتور در موفقیت XGBoost مقیاس‌پذیری آن در تمام سناریوهاست. محاسبات موازی و توزیع‌شده منجر به یادگیری ۱۰ برابر سریع‌تر مدل نسبت به راه‌حل‌های رایج موجود روی یک ماشین می‌شود [۵].

الگوریتم XGBoost به دلیل ترکیب روش‌های خلاقانه و بهینه‌سازی‌های موثر بسیار مقیاس‌پذیر است که تعدادی از آنها عبارتند از:

- یادگیری درختی موازی برای داده‌های پراکنده با پیچیدگی خطی از مرتبه زمانی تعداد داده‌های موجود
- طرح چندک وزن‌دار در یادگیری درختی تقریبی
- محاسبات خارج از حافظه اصلی[3] برای پردازش صدها میلیون داده روی یک کامپیوتر
- ساختار بلاکی آگاه از حافظه پنهان برای یادگیری درختی خارج از حافظه اصلی همراه با تکنیک‌های فشرده‌سازی داده، موازی‌سازی و خردکردن بلاک‌های داده
- نهایتاً همه این ویژگی‌ها در یک سیستم کامل ابتدا تا انتها[4] مجتمع می‌شوند و یک راه‌حل نوین برای مسائل دنیای واقعی ارائه می‌دهد.

با توجه به ویژگی‌های برجسته مدل XGBoost، در این پژوهش برای نخستین بار در حل مسئله اسپم ، این مدل مورد استفاده قرار گرفته است.

این الگوریتم یک روش یادگیری ماشین است که بر اساس یک مجموعه آموزشی شامل بردار ویژگی‌های صفحات اسپم و برچسب کلاس هر صفحه، آموزش می‌بیند و سپس با اجرا روی مجموعه آزمایشی و مقایسه نتایج حاصل از الگوریتم دسته‌بندی با برچسب‌های کلاس آن مجموعه، ارزیابی شده و دقت آن محاسبه می‌شود.

در این مقاله از مجموعه داده WEBSPAM-UK۲۰۰۷ برای آموزش و آزمایش مدل دسته‌بندی پیشنهادی استفاده شده است. این مجموعه داده در حال حاضر محبوب‌ترین و کامل‌ترین مجموعه داده در حوزه شناسایی صفحات اسپم است.

۴-۱-۲ معرفی مجموعه داده WEBSPAM-UK۲۰۰۷

مجموعه داده WEBSPAM-UK۲۰۰۷ در سال ۲۰۰۷ توسط کاستیلو [۴] با هدف استفاده در مطالعات تشخیص اسپم وب جمع‌آوری شده است. بخشی از این مجموعه توسط داوطلبان به طور دستی در سه کلاس نرمال، اسپم و دسته‌بندی‌نشده، برچسب‌گذاری شده و برچسب‌ها در دو مجموعه منتشر شده‌اند. این مجموعه داده شامل ۱۱۴۵۲۹ هاست است که ۶۴۷۹ هاست برچسب‌گذاری شده‌اند. مجموعه اول شامل دوسوم هاست وب ارزیابی‌شده برای آموزش و مجموعه دوم شامل یک‌سوم باقیمانده برای آزمایش مدل ارائه شده است.

شمای کلی داده‌های دو مجموعه آموزش و آزمایش مرورشده توسط ارزیابان در شکل‌های ۱ و ۲ رسم شده است.

کاستیلو و همکاران به جمع‌آوری مشخصات صفحات وب پراختند و برای هر صفحه وب در این مجموعه داده ۲۷۵ ویژگی مختلف در ۴ گروه ویژگی‌های بدیهی هر صفحه، ویژگی‌های محتوایی، ویژگی‌های پیوند و یک گروه ویژگی‌های حاصل از انجام محاسبات بر روی ویژگی‌های پیوند برای صفحات وب محاسبه کردند. این صفحات شامل صفحه اصلی هر

1. Extreme Gradient Boosting
2. Regression Tree
3. Out of Core
4. End to End

جدول ۱: لیست ویژگی‌های مورد استفاده در طبقه‌بندی پیشنهادی.

نام ویژگی	شرح ویژگی
HST_19	میزان بازیابی برای ۱۰۰ پرس‌وجوی برتر
HST_20	میزان بازیابی برای ۲۰۰ پرس‌وجوی برتر
Outdegree_hp	درجه خروجی هاست
Pagerank_hp	امتیاز PageRank هاست
Truncatedagerank_1_hp	امتیاز TruncatedPageRank در فاصله همسایگی ۱
Truncatedagerank_2_hp	امتیاز TruncatedPageRank در فاصله همسایگی ۲
Truncatedagerank_3_hp	امتیاز TruncatedPageRank در فاصله همسایگی ۳
Truncatedagerank_4_hp	امتیاز TruncatedPageRank در فاصله همسایگی ۴
L_outdegree_hp	لگاریتم درجه خروجی هاست
L_Pagerank_hp	لگاریتم امتیاز PageRank هاست
L_Truncatedagerank_1_hp	لگاریتم امتیاز TruncatedPageRank در فاصله همسایگی ۱
L_Truncatedagerank_2_hp	لگاریتم امتیاز TruncatedPageRank در فاصله همسایگی ۲
L_Truncatedagerank_3_hp	لگاریتم امتیاز TruncatedPageRank در فاصله همسایگی ۳
L_Truncatedagerank_4_hp	لگاریتم امتیاز TruncatedPageRank در فاصله همسایگی ۴

هاست و صفحه با بالاترین امتیاز PageRank در آن هاست می‌باشد. مجموعه داده ورودی ما شامل ۱۴ ویژگی انتخاب‌شده از مجموعه ویژگی WEBSPAM-UK۲۰۰۷ برای صفحه اصلی هر هاست می‌باشد. لیست ویژگی‌های استفاده‌شده در این مقاله در جدول ۱ آمده است. ما در این پژوهش از مجموعه ویژگی پیشنهادشده توسط اصدقی و سلیمانی حاصل از روش انتخاب ویژگی smart_BT معرفی‌شده در [۸] برای انتخاب ویژگی‌ها استفاده می‌کنیم. در این روش ۱۴ ویژگی هر صفحه انتخاب شده است؛ در صورتی که در اکثر تحقیقات مشابه از تعداد بسیار بیشتری ویژگی برای شناسایی صفحات اسپم استفاده شده است. در [۷] نیز از این مجموعه ویژگی استفاده شده است.

در این مرحله ابتدا تمام داده‌ها با برچسب «دسته‌بندی‌نشده» از مجموعه داده حذف شدند؛ سپس عملیات پیش‌پردازش شامل نرمال‌سازی داده‌ها برای داده‌های باقیمانده انجام شد. پس از آموزش مدل روی مجموعه آموزشی و آزمایش آن بر روی مجموعه داده تست، صفحات اسپم با دقت ۹۴٫۲۷٪ شناسایی شدند که در مقایسه با بهترین روش‌های موجود، بالاترین دقت است.

چالش مهم دیگری که در مسئله وب وجود دارد، این است که طبیعتاً تعداد صفحات خوب نسبت به صفحات اسپم بسیار بیشتر است؛ هرچند تمرکز سازندگان اسپم وب حضور در اولین نتایج جستجو است و کیفیت جستجو را به شدت تحت تأثیر قرار می‌دهند. در همه مجموعه داده‌های موجود در این حوزه نیز نسبت اسپم به نرمال بسیار کم است؛ اما نسبت آماری این صفحات کمتر از صفحات نرمال می‌باشد. در مجموعه داده WEBSPAM-UK۲۰۰۷ همان طور که در شکل‌های ۱ و ۲ مشخص است، این نسبت حدود ۵٪ است.

عدم توازن کلاس‌ها منجر به تبعیض در دسته‌بندی به نفع کلاس اکثریت می‌شود؛ در نتیجه سیستم دسته‌بندی صفحات اسپم را با دقت کمتری یاد می‌گیرد. پس «نرخ اسپم‌هایی که امتیاز نرمال می‌گیرند» بیشتر از «نرخ نرمال‌هایی که امتیاز اسپم می‌گیرند» خواهد شد ($FP > FN$). همچنین معیارهای ارزیابی سنتی مانند دقت می‌توانند در داده‌های نامتوازن به ارزیابی‌های گمراه‌کننده منجر شوند؛ یعنی یک مدل می‌تواند با پیش‌بینی کلاس اکثریت، دقت بالا داشته باشد؛ همان طور که دقت پایه در این مسئله، درصد قابل توجهی است. در صورتی که در بسیاری از کاربردهای واقعی، چنین پیش‌بینی‌هایی ارزش زیادی ندارند. مدل‌های آموزش‌دیده بر روی داده‌های نامتوازن ممکن است دچار مشکل در تشخیص وقوع رویدادهای کمیاب یا نادر شوند و در کاربردهایی مانند تشخیص صفحات اسپم به خسارت‌های مالی قابل توجهی برای شرکت موتور جستجو و نارضایتی کاربران منجر شود. با ایجاد توازن در کلاس‌ها، این تبعیض برطرف خواهد شد.

۲-٤ دسته‌بندی XGSspam

در دومین راهکار ارائه‌شده در این پژوهش، ابتدا داده‌های مجموعه داده WEBSPAM-UK۲۰۰۷ متوازن‌سازی شدند و سپس داده متوازن با الگوریتم XGBoost دسته‌بندی شد.

قابلیت هر تکنیک متوازن‌سازی نه فقط به اصل عملیاتی آن، یعنی نحوه متعادل‌کردن کلاس‌های متغیر هدف، بلکه به نسبت عدم تعادل بین کلاس‌های متغیر هدف نیز بستگی دارد. در آن مواردی که عدم تعادل کلاس‌های یک مجموعه داده بسیار شدید می‌باشد، تکنیک‌های متوازن‌سازی بیش‌نمونه‌برداری دارای نتایج بهتری بوده‌اند [۲۳]. با توجه به شدیدبودن نسبت عدم تعادل کلاس‌های اسپم و نرمال در مجموعه داده WEBSPAM-UK۲۰۰۷ مطابق با شکل‌های ۱ و ۲، تکنیک بیش‌نمونه‌برداری SMOTE پیشنهاد می‌شود. ضمن اینکه در مسایل طبقه‌بندی داده‌های نامتعادل، SMOTE می‌تواند مشکل اضافه برازش را کاهش دهد و عملکرد طبقه‌بندی‌کننده را بهبود بخشد [۲۴]. این روش با ترکیب دو تکنیک کارا، موفق به دسته‌بندی صفحات اسپم با دقت بسیار خوب ۹۵٫۴۴٪ شد.

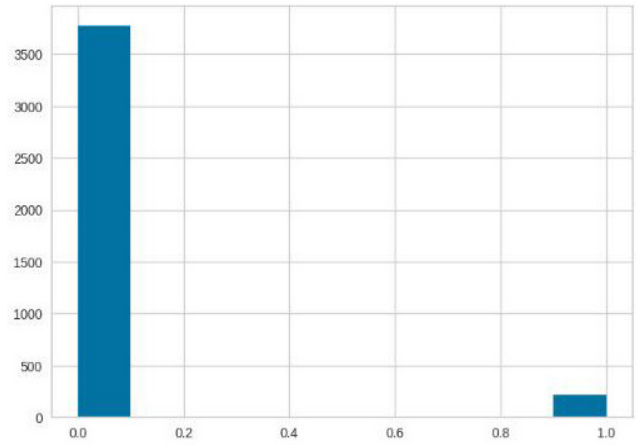
الگوریتم SMOTE یک تکنیک ایجاد بیش‌نمونه است که در آن کلاس اقلیت با ایجاد نمونه‌های ساختگی، بیش‌نمونه‌برداری می‌شود. این رویکرد از تکنیکی موفق در تشخیص کاراکترهای دست‌نویس الهام گرفته شده که در آن داده‌های اضافی را با ایجاد تغییراتی روی داده‌های واقعی ایجاد کردند؛ مثل چرخش یا کشش حروف.

SMOTE داده‌های ساختگی را با کار در فضای ویژگی‌ها به جای فضای داده، ایجاد می‌کند. با توجه به تعداد نمونه ساختگی مورد نیاز، به طور تصادفی k تا از نزدیک‌ترین همسایه‌ها انتخاب می‌شوند؛ سپس ویژگی‌های نمونه جدید کلاس اقلیت بر اساس ویژگی‌های همسایگان مربوطه محاسبه می‌شود. مثلاً اگر میزان داده اضافی مورد نیاز ۲۰۰٪ باشد، از بین نزدیک‌ترین همسایگان ۲ همسایه انتخاب شده و در راستای هر کدام یک نمونه جدید ایجاد می‌شود.

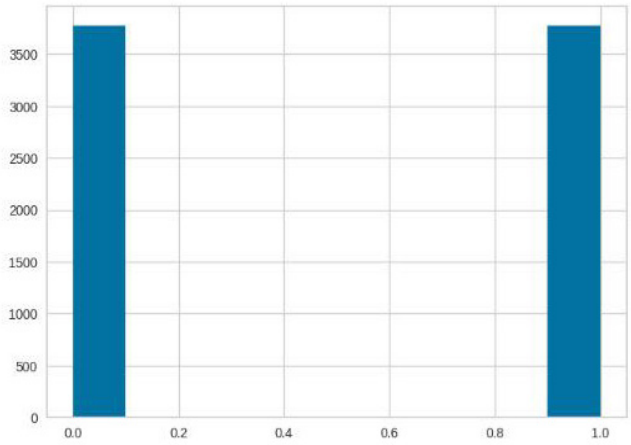
در SMOTE، نمونه‌های جدید با درون‌یابی خطی بین نمونه x_i که به طور دلخواه از کلاس اقلیت انتخاب شده و هر x_j از k نزدیک‌ترین همسایه‌های آن نمونه تولید می‌شود. روش سنتز نمونه‌های کلاس اقلیت در (۳) آمده است

$$X_{new} = x_i + rand(۰,۱) \times (x_i - x_j) \qquad (۳)$$

که در آن $rand(۰,۱)$ نشان‌دهنده یک عدد تصادفی انتخاب‌شده در بازه $(۰,۱)$ است.



شکل ۳: توزیع صفحات اسپم و نرمال در مجموعه آموزش قبل از متوازن‌سازی.



شکل ۴: توزیع صفحات اسپم و نرمال در مجموعه آموزش بعد از متوازن‌سازی.

جدول ۲: نتایج اجرای روش‌های پیشنهادی.

	XGspam	XGspam	XGSspam
تعداد ویژگی	۲۷۷	۱۴	۱۴
دقت	۰/۹۵۲۵	۰/۹۴۲۷	۰/۹۵۴۴
صحت کلاس ۰	۰/۹۶	۰/۹۵	۰/۹۵
بازخوانی کلاس ۰	۰/۹۹	۰/۹۹	۰/۹۶
صحت کلاس ۱	۰/۷۱	۰/۳۹	۰/۹۵
بازخوانی کلاس ۱	۰/۲۶	۰/۱۲	۰/۹۵
auc	۰/۶۲۷۵	۰/۵۵۴۳	۰/۹۵۴۴
زمان آموزش مدل	۴/۹۵ ثانیه	۰/۴۰ ثانیه	۰/۷۴ ثانیه

SMOTE نمونه‌های کلاس اقلیت مرتبط بیشتری را برای یادگیری فراهم می‌کند؛ بنابراین به یادگیرنده اجازه می‌دهد تا مناطق تصمیم‌گیری گسترده‌تری را ایجاد کند که منجر به پوشش بیشتر طبقه اقلیت و در نتیجه آموزش بهتر مدل‌های یادگیری ماشین می‌شود [۶]. هنگام طبقه‌بندی داده‌های نامتعادل، SMOTE می‌تواند مشکل اضافه برازش را کاهش دهد و عملکرد طبقه‌بندی‌کننده را بهبود بخشد.

پس از متوازن‌سازی مجموعه داده با روش SMOTE نسبت صفحات دو کلاس اسپم و نرمال در مجموعه داده برابر شد. همان طور که در شکل‌های ۱ و ۲ مشهود است، اکثریت صفحات وب در مجموعه داده WEBSPAM-UK۲۰۰۷ صفحات نرمال هستند و صفحات اسپم در هر دو مجموعه آموزش و آزمایش حدود ۵ درصد از کل صفحات آن مجموعه است. با توجه به درصد بسیار پایین اسپم در مجموعه داده، پس از حذف صفحات «دسته‌بندی‌نشده» از مجموعه آموزش و آزمایش، ابتدا داده‌های دو مجموعه ترکیب شدند و عملیات پیش‌پردازش بر روی کل داده‌ها انجام شد و سپس داده‌ها با استفاده از تکنیک SMOTE متوازن شدند؛ یعنی برای کلاس اقلیت، نمونه‌های جدیدی در همسایگی نمونه‌های موجود در این کلاس تولید شد. پس از متوازن‌سازی، مجموعه شامل ۷۵۵۲ هاست است. پس از اتمام مرحله پیش‌پردازش[1]، داده‌ها به صورت تصادفی جابجا می‌شوند و مجدداً به دو مجموعه آموزش و آزمایش تقسیم[2] شدند. مجموعه آزمایش شامل ۲۰٪ از این مجموعه داده و مجموعه آموزش شامل ۸۰٪ مجموعه داده می‌باشد. نمای توزیع داده‌ها مجموعه آموزش، قبل و بعد از متوازن‌سازی در شکل‌های ۳ و ۴ نشان داده شده است.

پس از اینکه توزیع صفحات اسپم و نرمال به حالت مطلوب رسید، مجدداً مجموعه داده متوازن با استفاده از مدل XGBoost دسته‌بندی شد. به طور قابل توجهی دقت دسته‌بندی افزایش یافت و به ۹۵/۴۴٪ رسید که دقت بسیار مطلوبی در شناسایی صفحات اسپم وب است.

۵- نتایج تجربی

در این بخش نتایج حاصل از دو روش پیشنهادی XGspam و XGSspam بررسی می‌شود. ابتدا به محاسبه دقت پایه در مسئله شناسایی اسپم وب بر روی مجموعه داده WEBSPAM-UK۲۰۰۷ می‌پردازیم. سپس نتایج متریک‌های متفاوت حاصل از اجرای روش‌های پیشنهادی را بررسی می‌کنیم و در پایان نتایج حاصل با روش‌های مشابه مقایسه شده است.

۵-۱ محاسبه دقت پایه

دقت پایه در حالتی که هیچ تلاشی برای طبقه‌بندی صفحات انجام نشود و همه صفحات مجموعه، نرمال فرض شوند، با محاسبه درصد نرخ صفحات نرمال به کل صفحات محاسبه شده است. دقت پایه برای تشخیص صحیح صفحات نرمال در مجموعه آموزش برابر با ۸۸/۳۲٪ و در مجموعه آزمایش برابر با ۸۷/۷۰٪ است. نکته قابل توجه این است که دقت بسیاری از الگوریتم‌های پیشین از دقت پایه کمتر می‌باشد.

۵-۲ نتایج حاصل از دو روش پیشنهادی

در این مقاله دو روش دسته‌بندی برای صفحات وب ارائه شده و گزارشی از نتایج اجرای این روش‌ها در جدول ۲ آمده است. در ستون اول نتایج حاصل از اجرای روش XGspam بر روی کل ویژگی‌های مجموعه داده و در ستون دوم نتایج اجرای روش XGspam بر روی مجموعه‌ای از ویژگی‌های انتخابی شامل ۱۴ ویژگی و نتایج حاصل از اجرای روش XGSspam بر روی همان مجموعه در ستون سوم آمده است. با مقایسه نتایج روش XGspam بر روی کل مجموعه داده و اجرای آن بر روی مجموعه داده کوچک‌تر می‌بینیم متریک‌های صحت و بازخوانی کلاس نرمال (۰) و دقت در دو روش تقریباً برابر هستند. اما صحت و بازخوانی کلاس اسپم (۱) در مجموعه داده بزرگ‌تر بسیار بهتر است که با توجه به حذف تعداد بسیار زیاد ویژگی‌ها به طوری که تعداد ویژگی‌های مورد استفاده در XGspam بر روی مجموعه داده اصلی، ۲۰ برابر بیشتر است،

1. Preprocess
2. Split

جدول ۳: مقایسه نتایج روش پیشنهادی با پژوهش‌های پیشین.

نام روش	مجموعه داده	ویژگی	الگوریتم	دقت
بچتی [۱۰]	Uk ۲۰۰۲	پیوند	Truncated PageRank	۸۰٪
تولاس [۹]	خزش‌شده	محتوایی	C۴.۵	۸۶٪٫۲
لیو [۱۶]	ساخته‌شده	محتوا	CNN	۹۱٪
لئو [۷]	Uk ۲۰۰۷	پیوند و محتوا	Random Forest	۹۳٪
XGspam	Uk ۲۰۰۷	پیوند و محتوا	XGBoost	۹۴٪
XGSspam	Uk ۲۰۰۷	پیوند و محتوا	XGBoost	۹۵٪

این نتیجه قابل انتظار و بلکه بهتر از حد انتظار است. اما به دلیل ابعاد بزرگ‌تر داده، زمان آموزش این مدل بیش از ۸ برابر طولانی‌تر است. با توجه به ابعاد وب، شناسایی صفحات اسپم در کارایی مدل بسیار مؤثر است.

نتایج حاصل از روش XGSspam نشان می‌دهند که هرچند دقت این روش تقریباً معادل روش XGspam است، اما صحت و بازخوانی کلاس اسپم و همچنین معیار auc بهبود قابل توجهی داشته است؛ به طوری که تنها در معیار بازخوانی ۸۳٪ افزایش داشته‌ایم، هرچند زمان آموزش مدل حدوداً ۲ برابر طولانی‌تر شده است.

۵-۳ مقایسه نتایج با تحقیقات مشابه

در جدول ۳ تعدادی از مطالعات مرتبط و نتایج آنها آورده شده است. از بین روش‌های بررسی‌گردیده، تنها روش لئو [۷] بر روی مجموعه داده WEBSPAM-UK۲۰۰۷ اجرا شده و نتایج آن قابل مقایسه با روش‌های پیشنهادی در این مقاله است. روش بچتی [۱۰] بر روی مجموعه داده uk ۲۰۰۲ اجرا شده و فقط از ویژگی‌های پیوند استفاده کرده است. بقیه مدل‌ها روی مجموعه داده‌های خزش‌شده و ساخته‌شده توسط نویسندگان اجرا شده‌اند. به دلیل تنوع متریک‌های ارائه‌شده در مقالات مختلف، در اینجا تنها موفق به مقایسه دقت روش‌ها با یکدیگر شدیم.

همان طور که در جدول مشهود است، استفاده از الگوریتم‌های یادگیری ماشین و آموزش آنها بر اساس ترکیبی از ویژگی‌های پیوند و محتوا نتایج بهتری را به دست می‌دهد. در این مقاله علاوه بر انتخاب الگوریتم سریع و دقیق دسته‌بندی XGBoost و بهره‌بردن از طیفی از ویژگی‌های پیوند و محتوا، با ارائه راهکاری برای چالش نامتوازن‌بودن تعداد صفحات اسپم و نرمال در مجموعه داده، موفق به بهبود چشمگیر صحت، بازخوانی، دقت، auc و سرعت شناسایی صفحات اسپم شده‌ایم.

۶- نتیجه‌گیری و کارهای آینده

کاربران موتورهای جستجو، معمولاً به نتایج اول جستجو توجه می‌کنند. حضور صفحات اسپم در بالای لیست نتایج جستجو، منجر به اتلاف زمان کاربران و در نتیجه نارضایتی کاربران و کاهش اعتبار موتور جستجو خواهد شد. به همین دلیل موتورهای جستجو معمولاً هزینه بسیاری را صرف شناسایی و حذف یا کاهش رتبه صفحات اسپم می‌کنند. با این وجود با توسعه الگوریتم‌های رتبه‌بندی جدید، اسپمرها نیز از تکنیک‌های متناسب جهت کسب رتبه بالاتر استفاده می‌کنند. این چرخه همچنان ادامه دارد و مسئله اسپم وب هرگز به طور کامل حل نشده است.

در این مقاله از بخشی از مجموعه داده WEBSPAM-UK۲۰۰۷ شامل ۱۴ ویژگی مبتنی بر محتوا و لینک استفاده شده و به دو روش دسته‌بندی انجام شده است. در روش اول به نام XGspam یک مدل دسته‌بندی با استفاده از تکنیک XGBoost ایجاد گردید که موفق به دسته‌بندی صفحات وب در دو کلاس نرمال و اسپم با دقت ۹۴٪ شد. در روش دوم دو تکنیک متوازن‌سازی و دسته‌بندی درختی با هم ترکیب شدند. پس از آموزش و آزمایش این مدل به دقت و بازخوانی ۹۵٪ رسید که دقتی بالاتر از روش‌های موجود است. در این پژوهش تنها ۱۴ ویژگی هر صفحه بر اساس پژوهش‌های موفق پیشین انتخاب و استفاده شده است؛ در نتیجه با محاسبات کمتر و با سرعت بیشتر به دقت خوبی دست یافته و با یک الگوریتم یادگیری ماشین در دسترس، ساده، سریع و عملی موفق به کسب دقت دسته‌بندی قابل توجهی در مقایسه با کارهای پیشین شده است.

استفاده از تکنیک‌های انتخاب ویژگی و سپس دسته‌بندی داده‌ها با مدل ترکیبی XGSspam، امیدبخش حصول به دقت بالاتر در کارهای آینده است. همچنین بهره‌برداری هم‌زمان از ویژگی‌های محتوا و پیوند و ساختار گراف وب می‌تواند منجر به نتایج دقیق‌تری در آینده شود. با توجه به ابعاد وب و رشد روزافزون آن، امکان تولید محدود داده‌های برچسب‌دار، بهره‌بردن از داده‌های بدون برچسب نیز می‌تواند در حل مسئله اسپم وب بسیار راهگشا باشد.

مراجع

[1] E. Convey, "Porn sneaks way back on web," *The Boston Herald*, vol. 28, 1996.

[2] M. De Kunder, *"he Size of the World Wide Web (The Internet),* https://www.worldwidewebsize.com, Retrived 2024.

[3] A. Shahzad, N. M. Nawi, M. Z. Rehman, and A. Khan, "An improved framework for content-and link-based web-spam detection: a combined approach," *Complexity*, vol. 2021, Article ID: 6625739, 18 pp., 2021.

[4] C. Castillo, *Web Spam Collections*, https://chato.cl/webspam/datasets/uk2007, Retrived 2024.

[5] T. Chen and C. Guestrin, "Xgboost: a scalable tree boosting system," in *Proc. ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, pp. 785-794, San Francisco, CA, USA, 13-17 Aug. 2016.

[6] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Artificial Intelligence Research*, vol. 16, no. 1, pp. 321-357, Jan. 2002.

[7] J. Liu, Y. Su, S. Lv, and C. Huang, "Detecting web spam based on novel features from web page source code," *Security and Communication Networks*, vol. 2020, Article ID: 6662166, 14 pp., 2020.

[8] F. Asdaghi and A. Soleimani, "An effective feature selection method for web spam detection," *Knowledge-Based Systems*, vol. 166, pp. 198-206, Feb. 2019.

[9] A. Ntoulas, M. Najork, M. Manasse, and D. Fetterly, "Detecting spam web pages through content analysis," in *Proc. World Wide Web*, pp. 83-92, Edinburgh, Scotland, 23-26 May 2006.

[10] 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. the WebKDD*, 10 pp., 2006.

[11] R. Baeza-Yates, P. Boldi, and C. Castillo, "Generalizing PageRank: damping functions for link-based ranking algorithms," in *Proc. Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pp. 308-315, Seattle, WA, USA, 6-11 Aug. 2006.

[12] M. Yu, J. Zhang, J. Wang, J. Gao, T. Xu, and R. Yu, "The research of spam web page detection method based on web page differentiation and concrete clusters centers," in *Proc. Int. Conf. on Wireless Algorithms, Systems, and Applications*, pp. 820-826, Tianjin, China, 20-22 Jun. 2018.

[13] J. J. Whang, Y. S. Jeong, I. Dhillon, S. Kang, and J. Lee, "Fast asynchronous antitrust rank for web spam detection," in *Proc. WSDM Workshop on Misinformation and Misbehavior Mining on the Web*, 4 pp., Marina Del Rey, CA, USA, 5-9 Feb. 2018.

[14] Z. Gyöngyi, H. Garcia-Molina, and J. Pedersen, "Combating web spam with trustrank," in *Proc.Very Large Data Bases*, vol. 30, pp. 576-587, Toronto, Canada, 31 Aug.-3 Sept. 2004.

[15] M. Sobek, Pr0-Google's Pagerank 0 Penalty, http://pr.efactory.de/e-pr0.shtml, Retrived 2024.

[16] D. Liu and J. Lee, "CNN based malicious website detection by invalidating multiple web spams," *IEEE Access*, vol. 8, pp. 97258-97266, 2020.
[17] X. Zhuang, Y. Zhu, Q. Peng, and F. Khurshid, "Using deep belief network to demote web spam," *Future Generation Computer Systems*, vol. 118, pp. 94-106, May 2021.
[18] C. Wei, Y. Liu, M. Zhang, S. Ma, L. Ru, and K. Zhang, "Fighting against web spam: a novel propagation method based on click-through data," in *Proc. Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pp. 395-404, Portland, ON, USA, 12-16 Aug. 2012.
[19] A. Heydari, M. A. Tavakoli, N. Salim, and Z. Heydari, "Detection of review spam: a survey," *Expert Systems with Applications*, vol. 42, no. 7, pp. 3634-3642, May 2015.
[20] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine," *Computer Networks and ISDN Systems*, vol. 30, no. 1-7, pp. 107-117, Apr. 1998.
[21] D. Sculley, *Kaggle: Your Machine Learning and Data Science Community*, https://www.kaggle.com, Retrived 2024.
[22] X. Ren, *Knowledge Dscovery in Data and Data-Mining*, https://kdd.org/, Retrieved 2024.
[23] T. Wongvorachan, S. He, and O. Bulut, "A comparison of undersampling, oversampling, and SMOTE methods for dealing with imbalanced classification in educational data mining," *Information*, vol. 14, no. 1, Article ID: 54, 2023.
[24] Y. Zhang, L. Deng, and B. Wei, "Imbalanced data classification based on improved random-SMOTE and feature standard deviation," *Mathematics*, vol. 12, no. 11, Article ID: 1709, 2024.

ریحانه رشیدپور تحصیلات خود را در مقطع کارشناسی مهندسی کامپیوتر گرایش نرم-افزار در سال ۱۳۸۸ از دانشگاه آزاد اسلامی شیراز و کارشناسی ارشد در سال ۱۳۹۰ از دانشگاه شهید بهشتی به پایان رسانده است و در سال ۱۴۰۳ از رساله دکتری خود در دانشگاه یزد دفاع کرده است و اکنون به خدمت معلمی در آموزش و پرورش با عنوان هنرآموز کامپیوتر مشغول می باشد. نام‌برده تجربه تدریس در دانشگاه‌های متعددی را دارد. زمینه‌های تحقیقاتی مورد علاقه ایشان عبارتند از: هوش مصنوعی، محاسبات گراف، معماری‌های نرم‌افزار و مهندسی نرم‌افزار.

علی‌محمد زارع بیدکی تحصیلات خود را در مقطع کارشناسی در سال ۱۳۷۸ از دانشگاه صنعتی اصفهان و مقاطع کارشناسی ارشد و دکتری کامپیوتر به‌ترتیب در سال‌های ۱۳۸۱ و ۱۳۸۸ از دانشکده فنی دانشگاه تهران به پایان رسانده است و هم‌اکنون عضو هیأت علمی دانشکده مهندسی کامپیوتر دانشگاه یزد می‌باشد. زمینه‌های تحقیقاتی مورد علاقه ایشان شامل بازیابی اطلاعات، موتورهای جستجو، رتبه‌بندی و پردازش زبان‌های طبیعی می‌باشد.