فصلنامه مدیریت اطلاعات و دانش‌شناسی
سال چهارم، شمارهٔ یک، بهار ۱۳۹۶ (۵۰-۴۱)

Quarterly Journal of Knowledge and Information Management
Vol 4, No 1, spring 2018 (41-50)

ارائه رویکردی به‌منظور شناسایی مجلات ربوده شده با استفاده از الگوریتم‌های درخت تصمیم

***منا اندوهگین‌شهری[۱]، بهرام امینی[۲]، محمد داورپناه جزی[۳]**

۱. دانشجوی کارشناسی ارشد مهندسی فناوری اطلاعات، موسسه آموزش عالی صنعتی فولاد، اصفهان، ایران.
۲. استادیار مهندسی کامپیوتر و فناوری اطلاعات، موسسه آموزش عالی صنعتی فولاد، اصفهان، ایران.
۳. استادیار مهندسی کامپیوتر و فناوری اطلاعات، موسسه آموزش عالی صنعتی فولاد، اصفهان، ایران.

دریافت (۱۳۹۶/۰۶/۰۵) پذیرش: (۱۳۹۶/۱۱/۰۵)

Presenting An Approach For Recognizing The Detecting Hijacked Journals by Using Decision Tree Algorithms

***Mona Andoohgin Shahri[1], Bahram Amini[2], Mohammad Davarpanah Jazi[3]**

1. MSc Student of Computer and Information Technology, Foulad Institute of Technology, Isfahan, Iran.
2. Assistant Professor of Computer and Information Technology, Foulad Institute of Technology, Isfahan, Iran.
3. Assistant Professor of Computer and Information Technology, Foulad Institute of Technology, Isfahan, Iran.

Received: (27/08/2017) Accepted: (25/1/2018)

چکیده

هدف: پژوهش حاضر با هدف تحلیل مجلات ربوده شده و ارائهٔ ویژگی‌هایی جدید در جهت به‌کارگیری آنها و کاهش کلاهبرداری اینترنتی در این زمینه انجام شد.

روش‌شناسی: برای دست‌یابی به این هدف، ابتدا براساس مطالعات انجام شده بر روی وب‌سایت مجلات قانونی و مجلات ربوده شده ویژگی‌های این نوع از حملات استخراج گشته و سپس مجموعهٔ آموزشی ایجاد می‌شود. تعداد رکوردهای داده‌ای گردآوری شده ۱۰۴ عدد است و با استفاده از ابزار داده‌کاوی WEKA مورد تحلیل و ارزیابی قرار گرفت.

یافته‌ها: داده‌های گردآوری شده در این پژوهش با استفاده از ابزار داده‌کاوی WEKA مورد تحلیل قرار گرفته است. الگوریتم‌های استفاده شده در این پژوهش الگوریتم‌های درخت تصمیم می‌باشند. الگوریتمی که نرخ خطای کمتری داشته باشد انتخاب می‌شود. یافته‌ها نشان می‌دهد که نرخ خطا در این روش حدود ۹ درصد بود.

بحث و نتیجه‌گیری: مطالعات پیشین حاکی از رشد تعداد ناشران جعلی و مجلات ربوده شده است. ازاین‌رو، در این مقاله به روشی جهت شناسایی این نوع مجلات پرداخته شده است. نتیجهٔ به‌دست آمده از داده‌های گردآوری شده با استفاده از درخت تصمیم و تحلیل آن نشان داده شده است.

واژه‌های کلیدی

مجلات ربوده شده، کلاهبرداری اینترنتی، مجلات معتبر، انتشار یافته‌های علمی.

Abstract

purpose: This study was conducted for the analysis of hijacked journal and presenting new features to apply them and reduce internet frauds.

Methodology: To reach this goal, First according to studies done on the websites the of authentic and hijacked journals, the features of this kind of attacks are extracted and then a training dataset is created. The number of collected data records is 104 were collected and analyzed using WEKA data mining tools. The results showed that the applied method has an error rate of 1 percent.

Findings: The collected data of this research was analyzed by Weka data mining software. The algorithms used for this survey were developed by Decision Tree Algorithm. An algorithm with lower error rate is selected. The results showed that the applied method has an error rate of 9 percent.

Conclusion: The previous studies showed an increase in the number of fake publishers and hijacked journals. This article is dealt with finding a way to identify these type of journals. The results obtained from the data collected using decision tree and the analysis thereof was shown.

Keywords

Hijacked Journal, Internet Fraud, Authentic Journal, Academic Ethics.

*نویسنده مسئول: منا اندوهگین شهری

*Corresponding Author: Mona Andoohgin Shahri
E-mail: andoohginmona@gmail.com

مقدمه

در اوایل سال ۲۰۱۲ جلالیان گزارش‌هایی درخصوص برخی از مجلاتی دریافت کرد که در تامسون رویترز دارای ضریب تأثیرند. این مجلات تمامی مقالات را در هر زمینه‌ای و با دریافت هزینه‌ای از نویسنده مقاله چاپ می‌کردند (جلالیان[1]، ۲۰۱۴) اصطلاح مجلات ربوده شده برای اولین بار در سال ۲۰۱۲ بعد از اینکه مجرمان سایبری در سال ۲۰۱۱ دامنهٔ منقضی شده "sciencerecord.com" را ثبت کردند مورد استفاده قرار گرفت. این کار جهت سوءاستفاده از نام دامنه به‌عنوان مکانی برای ترکیب سه مجله ربوده شده و هفت مجله جعلی بود. مجلات ربوده شده مورد استفاده science and nature، innova ciencia و science series data report می‌باشند (دادخواه و جلالیان[2]، ۲۰۱۵).

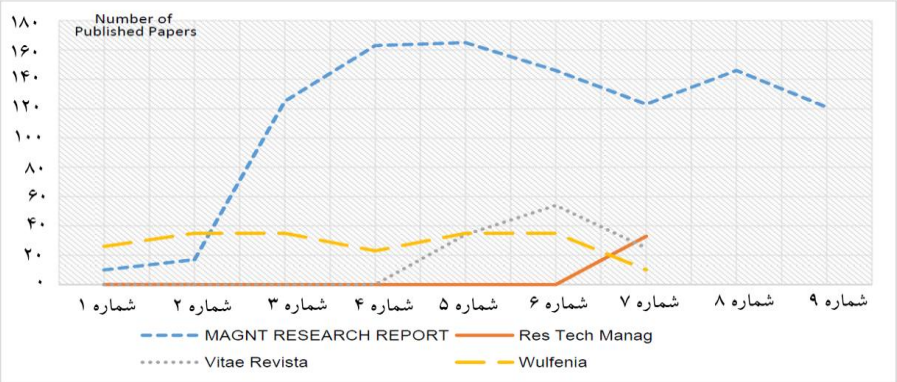
در سال ۲۰۱۳ باتلر از دو مجله اروپایی معروف گزارش داد که جاعلان هویت این مجله‌ها را سرقت و برای آنها وبسایت‌های تقلبی ایجاد کردند. این کار منجر به فریب‌خوردن صدها محقق و پرداخت هزینه به جاعلان شد (باتلر[3]، ۲۰۱۳). با توجه به گزارش باتلر مجلات معتبری که تنها دارای نسخه چاپی بوده و فاقد نسخه الکترونیکی‌اند بیشتر در معرض ربوده شدن توسط جاعلان قرار می‌گیرند. مشکلات بزرگ مجلات ربوده شده فقط محدود به سرقت منابع مالی قربانیان نمی‌شود. اضطراب و استرس قربانیان برای خراب‌شدن اعتبار آکادمیک و از دست‌دادن انگیزهٔ آنها برای تحقیقات علمی را نیز شامل می‌شود. همچنین مجلات قانونی که مورد ربایش قرار گرفته‌اند ممکن است اعتبار، نویسندگان و خوانندگان خود را از دست دهند. یافته‌های بررسی نشده مقالات در مجلات ربوده شده ممکن است اساس فرضیه‌ها و تحقیقات جدید شوند که در این صورت مجلات ربوده شده به اعتماد و اعتبار تحقیق منتشر شده و تحقیقات آیندهٔ حمله می‌کنند (جلالیان، ۲۰۱۴).

مجرمان اینترنتی با دزدیدن هویت مجلات قانونی و رفتن هزینه از فرهیختگان کلاهبرداری می‌کنند. این کار توسط یک چارچوب مناسب که شامل راه‌اندازی وبسایت، بازاریابی از طریق ایمیل و انتخاب قربانیان است، انجام میگیرد (لوکیچ و همکاران[4]، ۲۰۱۴).

مجرمان اینترنتی ایمیل‌های فراخوان را برای میلیون‌ها نفر از نویسندگان ارسال می‌کنند (جلالیان و محبوبی[5]، ۲۰۱۴).

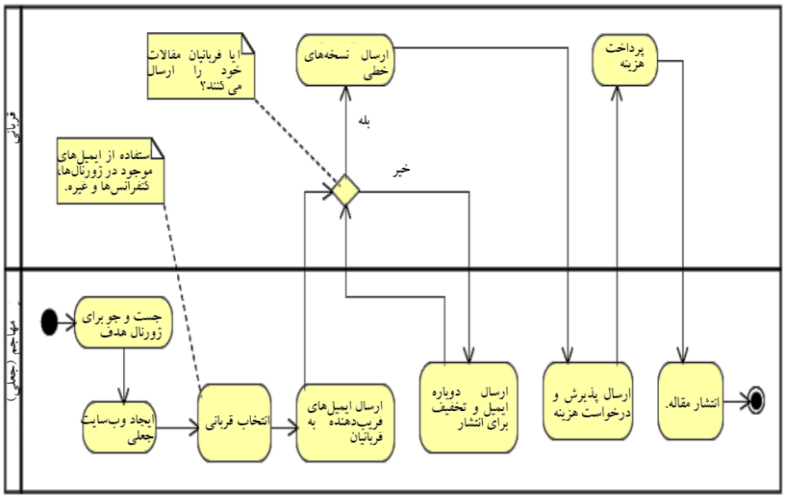
مجلات ربوده شده[6] یا ژورنال فیشینگ‌ها مجله‌هایی هستند که از عنوان و ISSN مجلات معتبر استفاده کرده و به فریب پژوهشگران می‌پردازند. جاعلان با ساخت وبسایت‌های جعلی و دریافت هزینه از محققانی که قصد چاپ یافته‌های خود را در مجلات نمایه شده دارند، کلاهبرداری می‌کنند (جلالیان و محبوبی، ۲۰۱۴). مجرمان اینترنتی مجلات غیرانگلیسی که از طریق ناشران کوچک اداره می‌شوند گزینهٔ خوبی برای مجرمان اینترنتی است. مجلات ربوده شده از مجلات معتبر تقلید کرده و به‌طور کلی ادعا می‌کنند که ضریب تأثیرشان را از تامسون رویترز کسب کرده‌اند. جاعل مجله به شخصی اطلاق می‌شود که مالک قانونی مجله نبوده و با ثبت دامنه‌ای برای مجله، مدعی می‌شود که وب‌سایت او وب‌سایت رسمی مجله است. به مجلاتی که هویتشان سرقت شده است، مجلات ربوده شده می‌گویند. این امکان وجود دارد که بعضی از مجلات ربوده شده دارای بیش از یک سایت جعلی باشند. یکی از این سایت‌ها معتبر بوده و توسط مالک قانونی آن اداره می‌شود که به آن مجلهٔ قانونی گویند. وب‌سایت‌های دیگر به‌عنوان وب‌سایت‌های جعلی در نظر گرفته می‌شوند که هیچ ارتباطی با مجله قانونی ندارند (جلالیان، ۲۰۱۵) با توجه به مقاله (دادخواه و بورچارت[7]، ۲۰۱۶) دامنهٔ مجلات جعلی در دو سال اخیر ثبت شده و رتبهٔ این صفحات در گوگل کمتر از چهار است. بنابراین، بسیاری از آنها قابل‌تشخیص هستند. ربودن مجلات در حال افزایش است. بیل لیستی از ۱۰۵ سایت مجله علمی را در سال ۲۰۱۶ منتشر کرد (بیل[8]، ۲۰۱۶). در مقاله (اندوهگین شهری و همکاران[9]، ۲۰۱۷) مجموعه داده و ویژگی‌هایی جدید برای شناسایی مجلات ربوده شده در نظر گرفته شد که با استفاده از الگوریتم های کلاس‌بندی این ویژگی‌ها ارزیابی شده و مورد تحلیل قرار گرفته‌اند. شکل ۱ تعداد مقالات چاپ شده در بعضی از محلات ربوده شده در هر شماره را نشان می‌دهد.

1. Jalalian
2. Jalalian & Dadkhah
3. Butler
4. Lukić & el al.
5. Jalalian & Mahboobi
6. Hijacked Journals
7. Dadkhah & Borchardt
8. Beall
9. Andoohgin Shahri & el al.



شکل ۱. تعداد مقالات چاپ شده در بعضی از مجلات ربوده شده در هر شماره
مأخذ: (دادخواه و همکاران، ۱۳۹۴)

مجله‌های ربوده شده همان‌طور که معرفی شدند، مجله‌هایی‌اند که با نام و اعتبار مجلات معتبر کار می‌کنند ولی هیچ ارتباطی با آن مجلات ندارند و همانند حملات فیشینگ صرفاً به‌دنبال انگیزه‌های مالی می‌باشند. با این تفاوت که در این نوع کلاهبرداری قربانی اطلاعات حساس خود را در اختیار مهاجم قرار نمی‌دهد بلکه به‌طور مستقیم منابع مالی را به مهاجم تحویل می‌دهد. شکل ۲ فرایند شکل‌گیری یک مجلهٔ ربوده شده و نحوهٔ فریب پژوهشگران به وسیلهٔ جاعلان را نشان می‌دهد.



شکل ۲. فرایند شکل‌گیری یک مجلهٔ ربوده شده و نحوهٔ فریب پژوهشگران به وسیلهٔ جاعلان
مأخذ: (دادخواه و همکاران، ۱۳۹۴)

روش پژوهش

این پژوهش از نوع کاربردی بوده و با استفاده از تکنیک‌های داده‌کاوی سعی در شناسایی مجلات ربوده شده دارد. برای داده‌های مورد نیاز در این پژوهش ابتدا براساس مطالعات انجام گرفته شده بر روی وب‌سایت مجلات قانونی و مجلات ربوده شده ویژگی‌های این نوع از حملات استخراج گردیده است. مقداردهی ویژگی‌ها براساس بررسی‌های انجام گرفته شده بر روی وب‌سایت‌های جمع‌آوری شده تعیین شده است. سپس مجموعه‌داده مورد نظر از وب‌سایت مجلات ربوده شده و مجلات قانونی استخراج شده و مجموعهٔ آموزشی ایجاد می‌شود و به فرمت مناسب برای ورود به WEKA تبدیل می‌شود. تعداد رکوردهای داده‌ای مورد بررسی در این پژوهش ۱۰۴ عدد است که در این بین ۵۹ رکورد مربوط به مجلات قانونی و ۴۵ داده مجله ربوده شده است. تجزیه و تحلیل داده‌ها با استفاده از نرم‌افزار WEKA انجام گرفته است. مجموعه داده ایجاد شده در نرم‌افزار WEKA وارد شده و سپس الگوریتم‌های درخت تصمیم بر روی داده‌ها اجرا شده و الگوریتمی که انتخاب می‌شود نرخ خطای کمتری داشته باشد.

پیشینه‌های پژوهش

در این قسمت ابتدا ویژگی‌های مورد استفاده جهت شناسایی مجلات ربوده شده معرفی شده سپس با استفاده از الگوریتم-های کلاس‌بندی، درخت تصمیمی جهت شناسایی مجلات ربوده شده ارائه و مورد تحلیل قرار می‌گیرد.

ویژگی‌های قابل‌استفاده جهت شناسایی مجلات ربوده شده

جدول ۱ ویژگی‌هایی که در این مقاله برای این نوع از حملات شناسایی شده‌اند را به همراه مقادیر قابل‌اندازه‌گیری برای هر ویژگی نشان می‌دهد. در این مقاله هشت ویژگی ارائه شده است. ویژگی توقف چاپ درزمینهٔ کشف مجلات جدید می‌باشد.

هفت ویژگی لینک‌های شکسته، تعداد مقالات چاپ شده در طول یک سال و همخوانی سرور و کشور (اندوهگین شهری و همکاران، ۲۰۱۷) رتبه دامنه در موتور جست‌وجو، طول عمر دامنه، کشورهای ورودی و دامنهٔ مقالات قابل‌قبول (دادخواه و همکاران، ۲۰۱۵) از پژوهش‌های قبلی گرفته شده‌اند. تعریف هریک از این ویژگی‌ها در ادامه آمده است.

جدول ۱. ویژگی‌های مورد استفاده در شناسایی مجلات ربوده شده.

نام ویژگی	نوع ویژگی	مقادیر
لینک‌های شکسته	پیوسته	اعداد طبیعی
تعداد شماره چاپ شده	پیوسته	اعداد طبیعی
همخوانی سرور و کشور مجله	منطقی	همخوانی سرور و کشور مجله =۱ عدم همخوانی سرور و کشور مجله =۰
رتبه دامنه در موتور جست‌وجو	منطقی	دارای رتبه =۱ فاقد رتبه =۰
طول عمر دامنه	منطقی	طول عمر کم =۰ طول عمر زیاد =۱
کشورهای ورودی	گسسته	بین ۱ تا ۴ کشور =H بین ۴ تا ۸ کشور =M بیش از کشور =L اطلاعاتی وجود ندارد =NA
دامنه مقالات قابل‌قبول	منطقی	دامنه عمومی =۰ دامنه اختصاصی =۱
توقف چاپ	منطقی	مجله قانونی توقف شده=۱ مجله فعال =۰

لینک‌های شکسته: به لینک‌هایی گفته می‌شود که کاربران را به صفحهٔ درستی هدایت نمی‌کنند و به مقصد نمی‌رسند. عواملی مانند بی‌دقتی در نوشتن لینک‌ها، حذف و یا جابجایی منابع و... عواملی هستند که در باوجود آمدن لینک‌های شکسته نقش ایفا می‌کنند. از آنجایی که سازندگان وب‌سایت‌های مجلات ربوده شده زمان زیادی را صرف ساخت این وب‌سایت‌ها نمی‌کنند بنابراین این سایت‌ها بیشتر در معرض این خطر قرار می‌گیرند. البته این ویژگی همیشه درست نخواهد بود چون ممکن است سازندگان وب‌سایت‌های قانونی هم نیز زمان زیادی را به رفع ایرادات موجود در سایت نپردازند

تعداد مقالات چاپ شده در یک سال: مجلات ربوده شده به‌دلیل عدم بررسی دقیق مقالات، نداشتن روند طولانی در پذیرش مقالات و داشتن انگیزه‌های مالی تعداد شماره‌های بیشتر و به‌طبع تعداد مقالات بیشتری در طول یک سال چاپ می‌کنند. بنابراین، با شمارش تعداد مقالات چاپ شده در طول یک سال و مقایسه آن با تعداد مقالات چاپ شده در مجلات معتبر می‌توان مجلات ربوده شده را شناسایی کرد .

همخوانی سرور و کشور مجله: طبق بررسی‌های انجام شده بر روی وب‌سایت‌های مجلات ربوده شده مشخص شد که در بیشتر آنها نام کشوری که سرور در آن قرار دارد با نام

کشوری که مجلات را منتشر می‌کند متمایز است. البته این ویژگی همیشه درست نخواهد بود چون ممکن است سرور و کشور مجله‌های معتبر به دلیل عدم توجه سردبیر و یا نیافتن سرور مناسب هم همخوانی نداشته نباشد اندوهگین شهری و همکاران، ۲۰۱۷).

دامنه در موتورهای جستجو: این ویژگی مربوط به دامنهٔ وب‌سایت مورد بررسی می‌باشد و بر پایه آن موتورهای جستجویی همچون گوگل وب‌گاه‌هایی که به هدف جستجوگر نزدیک‌ترند را در رده‌های بالاتری به دیگران قرار می‌دهد . معیار موتورهای جست‌وجو برای ارزیابی صفحات وب تجزیه و تحلیل لینک‌های ورودی به یک سایت است. رتبهٔ یک صفحه می‌تواند از ۰ تا ۱۰ باشد. از آنجایی که مجلات ربوده شده از وب‌سایت‌های اصلی کپی‌برداری می‌کنند. بنابراین، در موتورهای جستجو رتبهٔ بالایی به خود اختصاص نخواهند داد. البته این ویژگی همیشه درست نخواهد بود چون ممکن است مجله‌ای فاقد وب‌سایت باشد و موتورهای جست‌وجو وب‌سایت مجله ربوده شده را به‌عنوان وب‌سایت اصلی شناسایی کنند. این ویژگی به صورت یک متغیر بولی در نظر گرفته می‌شود چنانچه وب‌سایت مورد بررسی دارای رتبه در موتور جستجو باشد مقدار این متغیر یک و در غیر این صورت برابر صفر خواهد بود.

طول عمر دامنه: همخوانی طول عمر دامنه وب‌سایت متناسب با اولین شمارها و مقالات چاپ شده در مجله یکی از عواملی است که در وب‌سایت‌های متعلق به مجلات ربوده شده کمتر مورد توجه قرار می‌گیرد. با توجه به این موضوع و با استفاده از پایگاه Whois می‌توان مقدار این ویژگی را استخراج کرد. در این مقاله با اندکی تغییر مقدار این ویژگی به صورت یک متغیر بولی در نظر گرفته می‌شود چنانچه طول عمر دامنهٔ وب‌سایت متناسب با شماره مقالات چاپ شده باشد مقدار این متغیر یک و در غیر این صورت برابر صفر خواهد بود.

کشورهای ورودی به وب‌سایت مجلات: از آنجایی که وب‌سایت‌های فیشینگ دارای طول عمر کمتری نسبت به وب‌سایت‌های قانونی می‌باشند و محققین آشنایی چندانی با این وب‌سایت‌ها ندارند. بنابراین، در طبقه‌بندی بازدیدکنندگان دارای رتبهٔ پایین‌تری قرار دارند. البته این ویژگی همیشه درست نخواهد بود چون ممکن است مجله‌ای فاقد وب‌سایت باشد و یا توقیف شده باشد در این صورت ممکن است در طبقه‌بندی بازدیدکنندگان رتبهٔ بالاتری به آن تعلق گیرد. با استناد به پایگاه Alexa می‌توان مقدار این ویژگی را استخراج کرد و در شناسایی مجلات ربوده شده به کار گرفت.

دامنه مقالات قابل‌قبول مجله: اغلب مجلات ربوده شده به‌گونه‌ای هستند که مقالات با موضوعات مختلف را مورد پذیرش قرار می‌دهند و کمتر مجلاتی که مربوط به دامنهٔ خاصی هستند مورد جعل قرار می‌گیرند. اکثر این مجلات یا دارای اسامی خاص هستند که دامنهٔ موضوعی را مشخص نمی‌کنند و یا عنوان آنها به گونه است که محدوده وسیعی از فیلدهای تحقیقاتی را پوشش می‌دهد (دادخواه و همکاران، ۲۰۱۵).

توقف چاپ: طبق بررسی‌های انجام شده بر روی وب‌سایت‌های مجلات قانونی و ربوده شده مشخص گردید مجلات قانونی که توقف چاپ شده‌اند، فرصتی جهت سوءاستفاده برای جاعلان گشته‌اند. جاعلان با ساخت سایتی شروع به سوءاستفاده از نام و اعتبار مجلاتی که توقف چاپ شده‌اند می‌کنند. آنها شروع به ادامه چاپ با استفاده از نام این‌گونه مجلات کرده و محققان را فریب می‌دهند. این ویژگی به صورت یک متغیر بولین در نظر گرفته می‌شود چنانچه وب‌سایت قانونی مورد بررسی توقف شده باشد و دارای وب‌سایت جعلی باشد مقدار آن یک و در غیر این صورت برابر صفر خواهد بود.

داده‌های مورد استفاده

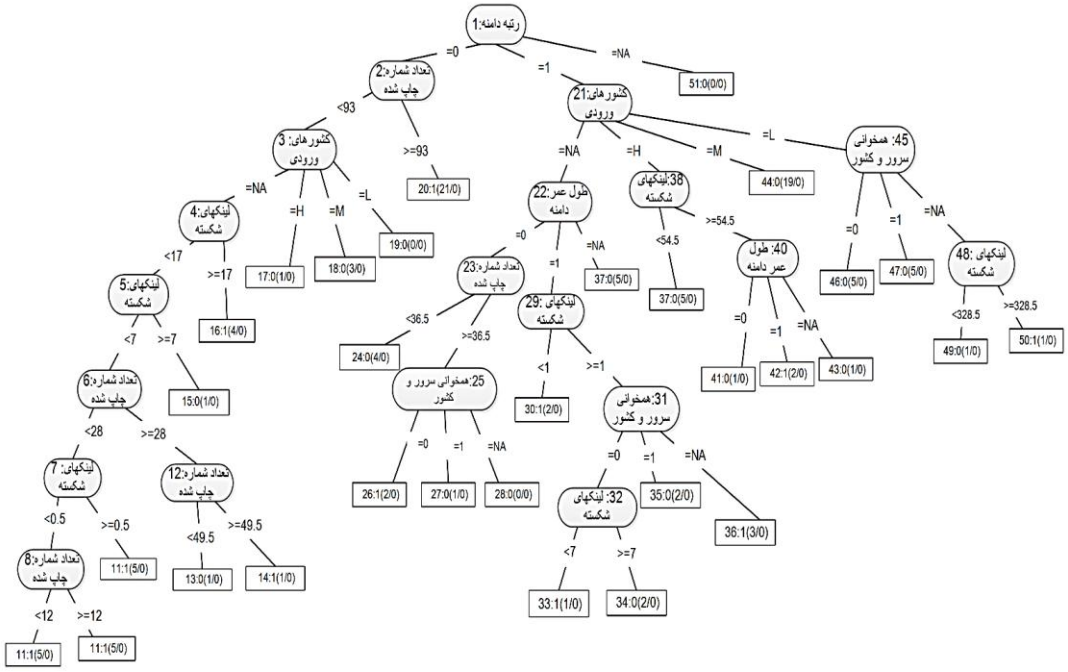
به‌دلیل جدیدبودن موضوع و در دسترس نبودن مجموعه داده در این زمینه داده‌های مورد نظر در این پژوهش توسط پژوهشگر ساخته شده که این روند شامل یافتن لیستی از مجلات معتبر و ربوده شده، بررسی سایت‌ها، در نظرگرفتن ویژگی، مقداردهی ویژگی‌ها، مرتب‌سازی داده‌ها و تبدیل مجموعه داده ایجاد شده به فرمت مناسب می‌باشد. مجموعه داده مورد نظر با بررسی لیستی از وب‌سایت‌هایی که توسط پژوهشگر جمع‌آوری شده مورد ارزیابی قرار گرفته است. در این روند ویژگی‌هایی جدید مطابق با بررسی سایت‌های مجلات قانونی و ربوده شده ارائه شده است. سپس هر مقدار ویژگی براساس بررسی‌های انجام گرفته شده بر روی وب‌سایت‌های جمع‌آوری شده تعیین شده و مجموعهٔ آموزشی ایجاد شده است.

استخراج درخت تصمیم از مجموعه داده‌ها

داده‌کاوی یکی از پیشرفت‌های اخیر در حوزهٔ کامپیوتر برای اکتشاف عمقی از داده‌ها است. پایگاه داده‌ها منبع غنی از اطلاعات پنهان هستند که با استفاده از طبقه‌بندی می‌توان مدلی جهت توصیف داده‌های موجود در آنها استخراج کرده و آنها را تحلیل کرد (هان و کامبر، ۲۰۰۶). بدین وسیله داده‌های با حجم بالا نیز بهتر فهمیده می‌شوند. طبقه‌بندی تلاش می‌کند تا رابطه‌ای میان متغیرهای مستقل با یک یا چند متغیر وابسته

کشف کند و درنهایت این رابطه با یک مدل نشان داده می‌شود. با کمک مدل می‌توان نمونه‌ها را به یکی از چندین طبقه‌ی تعریف شده منتسب و یا مقدار تعیین شده‌ای برای متغیر وابسته تعیین کرد. فرایند ساخت مدل یک فرایند دومرحله‌ای است. در مرحلهٔ اول که به مرحلهٔ یادگیری معروف است با استفاده از مجموعه داده‌های آموزشی که برچسب کلاس تمام نمونه‌های آن مشخص است مدل ساخته می‌شود. در مرحلهٔ دوم با کمک مجموعه داده‌های آزمایشی که در آن معمولاً برچسب کلاس‌ها نامعلوم است، مدل به‌دست آمده اعتبارسنجی می‌شود. در داده‌کاوی ابزارها و تکنیک‌های مختلفی براساس متدهای آماری استفاده می‌شود این متدها برای استخراج اطلاعات مفید بسیار مؤثر هستند. تکنیک‌های طبقه‌بندی را می‌توان به بخش‌هایی همچون قانون بیز، منطق فازی، الگوریتم ژنتیک، SVM، رگرسیون‌ها، طبقه‌بندی براساس تشابه، طبقه‌بندی مبتنی‌بر یافتن شروط و درخت تصمیم تقسیم کرد.

درخت تصمیم: این روش در داده‌کاوی مدلی جهت نمایش طبقه‌بندها و رگرسیون‌ها است. این درخت از تعدادی شاخه و گره تشکیل شده و برگ‌ها نمایانگر کلاس‌ها می‌باشند. در گره‌های دیگر به غیر از گره‌های برگ با توجه به یک یا چند متغیر هدف تصمیم‌گیری صورت می‌گیرد.



شکل ۳. درخت الگوریتم RandomTree

الگوریتم های DecisionStump، J48، LMT، RandomTree و REP Tree بر روی داده‌ها وارد شده و الگوریتمی که نرخ خطای کمتری داشته باشد انتخاب می‌شود. جدول ۲ نرخ خطای الگوریتم‌های گفته شده را نشان می‌دهند.

با توجه به نتایج به‌دست آمده الگوریتم Random Tree نرخ خطای کمتری دارد. شکل ۳ درخت این الگوریتم را نشان می‌دهد. در ادامه به بررسی درخت ارائه شده پرداخته می‌شود.

جدول ۲. نرخ خطای الگوریتم‌های اعمال شده

الگوریتم	نرخ خطا
DecisionStump	%۱۷.۳۰۷۷
J48	%۸.۶۵۳۸
LMT	%۷.۶۹۲۳
RandomTree	%۰.۹۶۱۵
REPTree	%۱۳.۴۶۱۵

در گرهٔ اول دامنه مجله مورد بررسی قرار می‌گیرد، اگر دامنه رتبه داشته باشد ۱ نداشته باشد ۰ و اگر نامشخص باشد NA است. فرض می‌شود دامنهٔ رتبه نداشته باشد در این صورت به شاخه سمت چپ وارد شده و تعداد مقالات چاپ شده در گره دوم بررسی می‌شود اگر این مقدار بیشتر و یا مساوی ۹۳ باشد مجله ربوده شده است. در غیر این صورت اگر کوچک‌تر از ۹۳ باشد کشورهای ورودی به سایت مجله در گره سوم بررسی می‌شود. اگر کشورهای ورودی تعدادشان M، H و L باشد مجله ربوده شده نیست در غیر این صورت اگر کشور ورودی نامشخص باشد لینک‌های شکسته در گرهٔ چهارم بررسی می‌شود. اگر تعداد لینک‌های شکسته بزرگ‌تر و یا مساوی ۱۷ باشد مجله ربوده شده است. در غیر این صورت دوبارهٔ لینک‌های شکسته در گرهٔ پنجم بررسی می‌شود. در صورتی که تعداد لینک‌های شکسته بزرگ‌تر و یا مساوی ۷ باشد مجله قانونی است. در غیر این صورت تعداد مقالات چاپ شده در گرهٔ ششم مورد بررسی قرار می‌گیرد. اگر تعداد مقالات چاپ شده بزرگ‌تر و یا مساوی ۲۸ باشد دوبارهٔ تعداد مقالات چاپ شده در گرهٔ ۱۲ مورد ارزیابی قرار می‌گیرد. درصورتی‌که تعداد مقالات چاپ شده بزرگ‌تر و یا مساوی ۵/۴۹ باشد مجله ربوده شده است در غیر این صورت مجله قانونی است. به گرهٔ شمارهٔ شش تعداد مقالات چاپ شده بازگشته اگر این مقدار کوچک‌تر از ۲۸ باشد تعداد لینک‌های شکسته در گرهٔ هفت مورد بررسی قرار می‌گیرند. در صورتی که این مقدار بزرگ‌تر و یا مساوی ۵/۰ باشد مجله ربوده شده است اما اگر این مقدار از ۵/۰ کوچک‌تر باشد تعداد مقالات چاپ شده در گرهٔ هشت مورد ارزیابی قرار می‌گیرد. اگر این مقدار بزرگ‌تر و یا مساوی ۱۲ باشد مجله ربوده شده و در غیر این صورت مجله قانونی می‌باشد.

به گرهٔ اول بازگشته، در صورتی که دامنهٔ مجلهٔ رتبه داشته باشد وارد شاخه مساوی یک شده و تعداد مقالات چاپ شده در گرهٔ ۲۱ مورد بررسی قرار می‌گیرد. در صورتی که کشورهای ورودی تعدادشان M باشد مجله ربوده شده نیست. اما زمانی که تعداد کشورهای ورودی به مجله نامشخص باشد طول عمر دامنه در گرهٔ ۲۲ مورد ارزیابی قرار می‌گیرد. اگر دامنهٔ مجله رتبه نداشته باشد تعداد شمارهٔ چاپ مقاله در گرهٔ ۲۳ مورد بررسی قرار می‌گیرد. درصورتی‌که این مقدار کمتر از ۳۶/۵ باشد مجله قانونی است. اما در زمانی که تعداد شمارهٔ چاپ مقاله بزرگ‌تر و یا مساوی ۵. ۳۶ باشد هم‌خوانی سرور و کشور مجله در گرهٔ ۲۵ بررسی خواهد شد اگر سرور و کشور مجله هم‌خوانی نداشته باشند مجله ربوده شده است. اما در صورتی که سرور و کشور مجله هم‌خوانی داشته باشند و یا وضعیت آنها نامشخص باشد مجله قانونی است. به گرهٔ ۲۲ طول عمر دامنه بازگشته و آخرین حالت آن یعنی طول عمر دامنه مساوی یک مورد بررسی قرار می‌گیرد. در صورتی که طول عمر دامنه یک باشد، لینک‌های شکسته در گرهٔ ۲۹ بررسی می‌شود اگر این مقدار کوچک‌تر از ۱ باشد مجله ربوده شده و در غیر این صورت هم‌خوانی سرور و کشور مجله در گرهٔ ۳۱ مورد ارزیابی قرار می‌گیرد. اگر سرور و کشور مجله همخوانی نداشته باشند لینک‌های شکسته در گرهٔ ۳۲ مورد بررسی قرار می‌گیرد. اگر این مقدار کوچک‌تر از هفت باشد مجله ربوده شده و در غیر این صورت قانونی می‌باشد. به گرهٔ ۳۱ همخوانی مجله و سرور مجله بازگشته در صورتی که کشور و سرور مجلهٔ همخوانی داشته باشند مجله قانونی و در صورتی که نامشخص باشد مجله ربوده شده می‌باشد. به گرهٔ ۲۲ طول عمر دامنه بازمی‌گشته در صورتی که طول عمر دامنه نامشخص باشد مجله قانونی است. به گرهٔ ۲۱ کشورهای ورودی به مجلهٔ بازگشته اگر تعداد این کشورها H باشد لینک‌های شکسته در گره ۳۸ بررسی می‌شود درصورتی‌که این مقدار کمتر از ۵/۵۴ باشد مجله قانونی است در غیر این صورت طول عمر دامنه در گرهٔ ۴۰ بررسی می‌شود. اگر طول عمر دامنهٔ صفر و نامشخص باشد مجله قانونی در غیر این صورت مجله ربوده شده می‌باشد. به گرهٔ ۲۱ کشورهای ورودی به مجله بازگشته اگر تعداد کشورهای ورودی L باشد همخوانی سرور و کشور مجله در گرهٔ ۴۵ مورد ارزیابی قرار می‌گیرد. در صورت همخوانی و عدم همخوانی سرور و کشور مجله، مجلهٔ قانونی و در صورت نامشخص بودن سرور و کشور مجله لینک‌های شکسته در گرهٔ ۴۸ مورد بررسی قرار می‌گیرد. اگر این مقدار کوچک‌تر از ۵/۳۲۸ باشد مجله قانونی و در غیر این صورت مجله ربوده شده می‌باشد. به گرهٔ اول رتبه‌بندی دامنهٔ بازگشته درصورتی‌که وضعیت رتبهٔ دامنه نامشخص باشد مجله قانونی است.

استفاده از رویکرد لیست سیاه

در مقالهٔ (جلالیان و بیملر[1]، ۲۰۱۴) به سرورهای یکسان در بعضی از مجلات ربوده شده اشاره شده است. با بررسی‌های انجام گرفته شده در این زمینه لیستی از سرورهای یکسان در این مجلات ارائه شده است. در بعضی از مجلات ربوده شده، جاعلان از سرورهای یکسان استفاده می‌کنند. IP سرور مجلات ذکر شده در لیست سیاه قرار می‌گیرد و اگر ایمیلی با IP سرور این مجلات ارسال به ایمیل کاربران شود به لیست دورریختنی فرستاده می‌شود. بنابراین، اگر IP سرور مجله مورد بحث در لیست سیاه باشد آن مجله مشکوک به ربوده شدن است. جدول ۳ (پس از منابع) لیست سروهای یکسان در مجلات ربوده شده را نشان می‌دهد.

1. Jalalian & Bimler

یافته‌های پژوهش

۱۰۴دادهٔ گردآوری شده در این مقاله با استفاده از ابزار داده‌کاوی مورد تحلیل قرار گرفته است. الگوریتم‌های استفاده شده در این مقاله الگوریتم‌های درخت تصمیم ازجمله DecisionStump، J48، Random Tree، LMT و REP Tree می‌باشند که بر روی‌داده‌ها وارد شده و الگوریتمی که نرخ خطای کمتری داشته باشد انتخاب می‌شود. سپس درخت تصمیم الگوریتمی که نرخ خطای کمتری داشته باشد ترسیم و مورد ارزیابی قرار می‌گیرد. به‌منظور تست رویکرد پیشنهادی مجموعه‌داده‌ای حاوی ۱۱ مجله قانونی و ربوده شده، ایجاد شده است. مجموعه تست ایجاد شده به‌عنوان ورودی به الگوریتم‌ها در ابزار WEKA وارد شده و میزان تشخیص صحیح الگوریتم‌ها اندازه‌گیری می‌شود. نرخ خطای رویکرد ارائه شده در این مقاله حدود ۹٪ می‌باشد.

بحث و نتیجه‌گیری

مجلات ربوده شده یکی از موضوعاتی است که کمتر مورد توجه قرار گرفته و همین امر سبب از بین رفتن یافته‌های علمی محققان می‌شود. مطالعات پیشین حاکی از رشد تعداد ناشران جعلی و مجلات ربوده شده می‌باشد. ازاین‌رو، در این مقاله به ارائه رویکردی به‌منظور شناسایی مجلات ربوده شده با استفاده از الگوریتم‌های درخت تصمیم پرداخته شده است. بسیاری از دانشجویان برای تحصیل در مقاطع بالاتر و اساتید دانشگاهی برای گرفتن ارتقای رتبه در طول سال، یافته‌های علمی خود را به مجلات دارای چندین نمایه معتبر ارسال می‌کنند. ازاین‌رو، سالیانه تعداد بی‌شماری مقاله برای چاپ به مجلات معتبر ارسال می‌شود که تنها تعداد اندکی از آنها و در طول فرایندی طولانی‌مدت به چاپ خواهند رسید. در این میان کلاهبردان اینترنتی با مشاهده مشکلات موجود در این فرایند به فریب محققان پرداختند. آنها با ساخت وب‌سایت‌های جعلی هم‌نام با مجلات معتبر از محققان کلاهبرداری می‌کنند. در این راه نتایج تحقیقات محققان با انتخاب اشتباه مجله ربوده شده به جای مجله معتبر از بین می‌رود. همان‌طور که ذکر شد در این مقاله به معرفی و ارائه روشی جهت شناسایی مجلات ربوده شده پرداخته شده است، در ابتدا به معرفی مجلات ربوده شده و اهمیت بحث در این موضوع پرداخته شد. سپس ویژگی‌هایی جهت شناسایی این نوع از حملات ذکر شد که امکان شناسایی مجلات ربوده شده را فراهم می‌کنند. در انتها با گردآوری لیست مجلات ربوده شده و معتبر و ارائه ویژگی‌های قابل‌تأمل مجموعه‌داده‌ای ایجاد شد که با اعمال الگوریتم‌های درخت تصمیم بر روی آن، الگوریتمی که کمترین نرخ خطا را داشت انتخاب و مورد تحلیل قرار گرفت. طبق تحلیل به‌دست آمده شش ویژگی رتبهٔ دامنه، تعداد شمارهٔ مقالات چاپ شده، لینک های شکسته، طول عمر دامنه، همخوانی سرور و کشور مجله و کشورهای ورودی مجلات از اهمیت بیشتری برخوردارند. درنهایت با ایجاد مجموعه‌ای حاوی ۱۱ داده رویکرد ارائه شده مورد ارزیابی قرار گرفت و نرخ خطای آن حدود ۹٪ می‌باشد.

Referances

Andoohgin, S. h., DavarpanahJazi, M. Borchardt, G. & Dadkhah, M. (2017). Detecting Hijacked Journals by Using Classification Algorithms. *Science and engineering Ethics*, 1-14.

Beall, J. (2016b). Retrieved 3 May 2016 from https://scholarlyoa.com/other-pages/hijacked-journals/.

Butler, D. (2013). Sham journals scam authors. *Nature*, 495, 421-422.

Dadkhah, M. & Borchardt, G. (2016). Hijacked Journals: An Emerging Challenge for Scholarly Publishing. *Aesthetic Surgery Journal*, 36(6), 739-741.

Dadkhah, M., Obeidat, M. M., Jazi, M. D., Sutikno, T. & Riyadi, M. A. (2015). How Can We Identify Hijacked Journals?. *Bulletin of Electrical Engineering and Informatics*, 4(2), 83-87.

Dadkhah, M., Sutikno, T., DavarpanahJazi, M. & Stiawan, D. (2015). An Introduction to Journal Phishings and Their Detection Approach. *Telecommunication, Computing, Electronics and Control*, 13(2), 656-660.

Han, J. & Kamber, M. (2006). *Data Mining Concepts and Techniques*. Second Edition, Diane Cerra, United States of America.

Jalalian, M. & Bimler, D. (2014). Retrieved 23 July from http://www.cybernewsalerts.com/2014/07/the-scientific-journal-interciencia-has.html.

Jalalian, M. (2014a). Journal hijackers target science and open access. *Research information*. From http://www.researchinformation.info/news/news_ story.php?news_id=1660

Jalalian, M. & Dadkhah, M. (2015). The full story of 90 hijacked journals. *Geographica Pannonica*, 19(2), 73-87.

Jalalian, M. & Mahboobi, H. (2014). Hijacked

Journals and Predatory Publishers: Is There a Need to Re-Think How to Assess the Quality of Academic Research?. *Walailak J. Sci & Tech*, 11(5), 389-394.

Jalalian, M. (2015). Solutions for commandeered journals, debatable journals, and forged journals. *Contemporary Clinical Dentistry*, 6(3), 283-285.

Lukić, T., Blešić, I., Ivanović, B. L. Milošević, D. & Sakulski, D. (2014). Predatory and Fake Scientific Journals/Publishers – A Global Outbreak with Rising Trend. *Geographica Pannonica*, 18(3), 69-81.

The Rise of Predatory Publishing: How To Avoid Being Scammed. (2016). Weed Science, 64(4), 772-778.

جدول ۳. لیست IP سروهای یکسان.

IP	آدرس سایت مجله
166.62.73.64	http://brisjast.com/ http://www.joetsite.com/ http://www.ijarsite.com/
89.45.67.200	http://www.jsrad.org http://www.waliaj.com
96.127.158.26	http://www.revistas-academicas.com http://www.ciencia-e-tecnica.org http://sylwan.ibles.org
5.9.88.196	http://canuorgme.org http://www.acoreanajr.com/
198.143.170.171	http://www.hfsp-journal.org http://www.jokulljournal.com http://www.kasmerajournal.com http://www.multidisciplinarywulfenia.org
111.118.181.161	www.ijmsi.org http://www.iosrjournals.org