دانشگاه آزاد اسلامی-واحد علوم و تحقیقات

دانشکده ی مهندسی کامپیوتر

سمینار دوره کارشناسی ارشد مهندسی کامپیوتر–گرایش نرم افزار

بررسی روشهای هوشمند شناسایی و کشف سرقت ادبی در متون علمی

نگارنده:

هاجر بغم

استاد سمینار:

دکتر امید سجودی

استاد راهنما:

دکتر بابک کرسفی

زمستان ۹۲

تقدیم به:

پدر فداکارم

کوهی استوار و حامی من در تمام مراحل زندگی

مادر مهربانم

سنگ صبوری که الفبای زندگی به من آموخت

سپاسگزاری:

با سپاس فراوان از استاد راهنمای فرهیخته‌ام جناب آقای بابک کرسفی که در طول مدت انجام این سمینار از رهنمودهای علمی و اخلاقی ایشان بهره‌مند شدم و درگاه خداوند بزرگ را شاکرم که افتخار شاگردی ایشان را نصیبم نمود .

از استاد سمینار گرامی جناب آقای دکتر امید سجودی به خاطر رهنمودهای علمی و اخلاقی ارزنده‌شان بسیار سپاسگزارم.

چکیده

سرقت ادبی از نقطه نظر «دزدی مالکیت معنوی» قدمتی به میزان تاریخ فعالیت‌های هنری و تحقیقاتی انسان‌ها دارد. با این حال، دسترسی آسان به وب، پایگاه‌های داده ای بزرگ و به‌طورکلی ارتباطات راه دور، باعث تبدیل سرقت ادبی به یک مشکل بزرگ برای ناشران، محققین و موسسات آموزشی شده است.

سرقت ادبی یک مشکل روزافزون در مؤسسات آکادمیک می‌باشد؛ و معمولاً به صورت کپی برداری از اثر فردی دیگر (مثلاً از دانشجویان دیگر و یا منابع دیگری مانند کتاب‌های درسی)، و عدم اشاره به منبع مطالب تعریف می‌شود. وفور منابع آنلاین موجود، فاکتوری است که تأثیرات زیادی بر افزایش وقوع سرقت ادبی در محیط‌های آکادمیک می‌گذارد زیرا استفاده از آثار دیگران برای دانشجویان آسان‌تر شده است.

در این سمینار ما ابتدا در فصل ۱ به معرفی برخی از تحقیقات انجام شده، سوابق بررسی سرقت ادبی و تاریخچه آن می‌پردازیم. پس از آن تحقیقات انجام شده از نظر دسته بندی روش‌های شناسایی راتوضیح می‌دهیم. سپس میگوییم که سرقت ادبی چیست و چه اهمیتی دارد. دسته بندی های سرقت ادبی در انتهای این فصل می آیند. فصل دوم به سرقت ادبی درون زبانی می پردازد و در فصل سوم سرقت ادبی میان زبانی و روشهای آن را بررسی میکنیم. در فصول چهارم و پنجم بترتیب سرقت ادبی در صفحات وب و سرقت ادبی در میان کد های کامپیوتری، به همراه اهمیت و روشهای تشخیص آن می آیند. در نهایت یک نتیجه گیری کلی از بحث در فصل آخر ارائه میگردد.

فهرست مطالب

فهرست جداول IX

فهرست تصاویر X

فصل اول : مقدمه XI

۱-۱ - مقدمه ۱

۱-۲- تحقیقات مرتبط و تاریخچه شناسایی سرقت ادبی ۲

۱-۳- سرقت ادبی چیست ۳

۱-۴- اهمیت سرقت ادبی ۴

۱-۵- مقیاس مشکل ۶

۱-۶- دسته‌بندی‌های سرقت ادبی ۷

۱-۷- تحقیقات انجام شده از نظر دسته بندی روش‌های شناسایی ۹

۱-۷-۱- اثر انگشت ها ۹

۱-۷-۲- شباهت رشته‌ای ۹

۱-۷-۳- روش‌های مبتنی بر ساختار ۱۰

۱-۷-۴- روش‌های خوشه بندی ۱۱

۱-۷-۵- مدلهای نحوی ۱۱

۱-۷-۶- روش‌های چند زبانی ۱۲

۱-۷-۷- روش‌های معنایی (صرفی) ۱۲

۱-۸- خلاصه فصل اول ۱۳

فصل دوم : سرقت ادبی در زبان ها ۱۵

۲-۱- سرقت ادبی درون زبانی ۱۶

۲-۱-۱ روش‌های مبتنی بر گرامر ۱۶

۲-۱-۲- روش معنایی ۱۷

۲-۱-۳- روش ترکیبی گرامر-معنایی ۱۷

۲-۱-۴- تشخیص سرقت ادبی بر مبنای اطلاعات ساختاری ۲۰

۲-۱-۵- روش تجزیه مقدار منفرد ۲۴
۲-۱-۶- نقشه های خود-آرایی (کوهوننیس) ۲۵
۲-۱-۷- ماتریس فاصله یکپارچه ۲۶
۲-۱-۸- برچسب زنی نقش معنایی ۲۷
۲-۱-۹- درهم سازی جملات ۲۸
۲-۱-۱۰- مدل و نمونه گیری ۳۰
۲-۱-۱۱- متن تکراری ۳۱
۲-۲- سرقت ادبی برون-زبانی ۳۴
انواع روش‌های شناسایی سرقت ادبی برون زبانی ۳۵
۲-۲-۱- سیستم‌های مبتنی بر لغت ۳۵
۲-۲-۲- سیستم‌های مبتنی بر فرهنگ لغت ۳۵
۲-۲-۳- سیستم‌های مبتنی بر مجموعه‌های داده قابل مقایسه ۳۵
۲-۲-۴- سیستم‌های مبتنی بر مجموعه‌های داده موازی ۳۶
۲-۲-۵- سیستم‌های مبتنی بر ترجمه ماشینی ۳۶
۲-۳- توصیف برخی از روش‌های مورد استفاده برای تشخیص سرقت ادبی برون زبانی ۳۶
۲-۳-۱- ابراز فرت ۳۶
۲-۳-۲- طرح سیستم فرت ۳۷
۲-۳-۴ N-گرم ۳۸
۲-۵- خلاصه فصل دوم ۴۰
فصل سوم : سرقت ادبی در کدهای کامپیوتری ۴۳
۳-۱- سرقت ادبی در علوم کامپیوتر ۴۴
۳-۲- مطالب سرقت شده در کدها ۴۵
۳-۳- انواع سیستم‌های شناسایی سرقت ادبی ۴۷
۳-۴- ابزارهای شناسایی سرقت ادبی کد منبع ۴۷
۳-۴-۱- سیستم‌های مبتنی بر اثر انگشت (فینگرپرینت) ۴۷

۳-۴-۲- الگوریتم های انطباق رشته ۵۰
۳-۴-۳- الگوریتم های انطباقی پارامتری سازی شده ۵۴
۳-۴-۴- الگوریتم های مقایسهٔ درخت‌های تجزیه ۵۵
۳-۵- مرور کوتاهی بر سرقت ادبی در صفحات وب ۵۹
۳-۵-۱- تحقیقات انجام‌شده ۶۰
خلاصه فصل سوم ۶۳
فصل چهارم : نتیجه گیری و چالش ها ۶۶
۴-۱- نتیجه گیری ۶۷
۴-۲- چالش ها ۶۷
منابع ۶۹
واژه نامه فارسی به انگلیسی ۷۱

فهرست جداول

جدول ۱: مثال [۱۲] ۳۰

جدول ۲: فهرست تری-گرم ها برای متن ۱ [۱۸] ۳۹

جدول ۳: فهرست تری-گرم ها برای متن ۲ [۱۸] ۳۹

جدول ۴: نتایج مفید توصیف شده. ستون Passage شامل متن ها است، N تعداد کلی تری-گرم ها در متن است، M تعداد تری-گرم های منطبق می‌باشد، و R ضریب تشابه است. [۱۸] ۴۰

جدول ۵ : مزایا ومعایب روش های تشخیص سرقت ادبی درون زبانی [۱۰] ۴۲

جدول ۷ : مزایا و معایب روش های تشخیص سرقت ادبی در کد نویسی ۶۵

فهرست تصاویر

تصویر ۱: نمونه‌ای از یک مورد سرقت ادبی و نمایش n-گرم کلمات کلیدی (۸- گرم مشترک متون با خط‌ها نشان داده شده‌اند) ۲۱

تصویر ۲: الگوریتم مربوط به تشخیص مرزهای متنی ۲۳

تصویر ۳: تجزیه مقدار منفرد کاهش یافته به k ۲۵

تصویر ۴: ساختار روش SRL ۲۸

تصویر ۵: طبقه بندی انواع سرقت ادبی متنی، به همراه روش‌های شناسایی آن‌ها ۳۴

تصویر ۶: سناریو ها و پاسخ‌ها ۴۶

فصل اول : مقدمه

۱-۱ – مقدمه

سرقت ادبی در مؤسسات آکادمیک معمولاً به صورت کپی برداری از اثر فردی دیگر (یعنی دانشجویان یا منابع دیگری مثل کتاب‌ها و مقالات) و عدم اشاره به نام منبع اثر (یعنی ارائه دهنده مطالب اصلی) تعریف می‌شود. سرقت ادبی بدون توجه به اینکه عمدی باشد یا غیرعمدی، جرم محسوب می‌شود.

هاناباس سرقت ادبی را به صورت «استفاده غیر مجاز و یا تقلید نزدیک از ایده‌ها و زبان/بیان فردی دیگر» تعریف می‌کند [۲].در حوزه آثار آکادمیک، سرقت ادبی می‌تواند دامنه‌ای از بیان چند جمله از یک اثر بدون اشاره به نویسنده یا گوینده آن تا کپی برداری از کل سند را شامل شود. سرقت ادبی یک جرم آکادمیک است و یک جرم حقوقی محسوب نمی‌شود، و به صورت قوانین و مقررات سازمانی مطرح می‌گردد. بدین ترتیب، مفهوم سرقت ادبی در میان سازمان‌ها و مؤسسات بر مبنای قوانین و مقررات آن‌ها متفاوت است. همه دانشگاه‌ها، سرقت ادبی را شکلی از کلاهبرداری یا سوءرفتار آکادمیک می‌دانند، اما قوانین و مقررات آن‌ها برای کنترل موارد مذنون به سرقت ادبی متفاوت است و جرایم اعمال شده بر این کلاهبرداری مبتنی بر فاکتورهایی مانند شدت جرم و اقرار فرد متخلف به جرم خود می‌باشد. در میان مؤسسات آکادمیک، جرایم متفاوتی اعمال می‌شود، که می‌تواند از نمره صفر در درس مرتبط با اثر سرقت شده، معدوم نمودن اثر، و یا در موارد جدی‌تر، اخراج از مؤسسه را در بر بگیرد.

لازم به ذکر فلینت ، تحقیقی در میان ۲۶ فرد آکادمیک انجام داد تا مفهوم سرقت ادبی در میان دانشجویان را از آن‌ها جویا شود[۲]. افرادی که در تحقیقات و حضور پیدا کردند در دپارتمان ها و دانشکده‌های مختلف یک دانشگاه کار می‌کردند. او دریافت که کارمندان دارای «تعریف شخصی» خود از سرقت ادبی هستند. تعاریف موردنظر کارمندان تحت تأثیر تجارب پیشین در موارد مرتبط با سرقت ادبی در آموزش عالی بود و همچنین مشخص شد که این تعاریف با سیاست‌های سازمان مورد نظر در خصوص سرقت ادبی همخوانی ندارد. این ناهماهنگی ها در تعریف سرقت ادبی در میان افراد آکادمیکی که در یک مؤسسه کار می‌کنند، نشان‌دهنده ناهماهنگی در دنبال کردن سیاست دانشگاه در هنگام شناسایی موارد سرقت ادبی است. این

مسئله می‌تواند به کار برد ناهماهنگ سیاست‌ها در میان افراد آکادمیک و همچنین اعمال جرایم/واکنش‌های متفاوت در برابر این موارد گردد. همچنین این مسئله می‌تواند باعث سردرگمی در میان دانشجویان پیرامون موضوع شود که سرقت ادبی در چه موردی اعمال می‌شود و در چه موردی که اعمال نمی‌شود.

سیستم‌های کامپیوتری و اتوماتیک تشخیص سرقت ادبی برای شناسایی این موارد در آثار مختلف به وجود آمده‌اند، و تأثیرگذاری این سیستم‌ها متکی بر نوع سرقت ادبی است که برای شناسایی آن به کار گرفته می‌شوند. این سیستم‌ها مزایای بسیاری از نظر صرفه‌جویی در زمان و فعالیت‌های افراد آکادمیک برای فرایند تشخیص دارم تشخیص کامپیوتری سرقت ادبی به این دلیل بر دو دهه گذشته مورد توجه افراد آکادمیک و محققین قرار گرفته است که استفاده از این ابزارها می‌تواند بار کاری آکادمیک را به واسطه اتوماتیک ساختن فرایندهای مقایسه و ارائه مطالب مشابه به طور قابل توجهی کاهش دهد.

۱-۲- تحقیقات مرتبط و تاریخچه شناسایی سرقت ادبی

تحقیقات اولیه در مورد ارزیابی سیستم‌های شناسایی سرقت ادبی اساساً بر روی توصیف نقاط قوت و ضعف سیستم‌های شناسایی متمرکز بودند توسط کلاف،لنکستر و کالوین، [۱].

استفاده از ابزارهای کامپیوتری شناسایی سرقت ادبی همچنین با یک‌سری مسائل حقوقی و اخلاقی نیز مرتبط هستند طبق نظریه فاستر، این مسائل هم به واسطه نقایص تکنولوژیکی الگوریتم های شناسایی سرقت ادبی (به‌عنوان مثال یک سیستم ممکن است تحقیقات یک دانشجو را به اشتباه سرقت ادبی تشخیص دهد) و هم به واسطه سوءتفاهم در خصوص نقش نرم‌افزارهای شناسایی سرقت ادبی در فرایندهای آموزشی پدید می‌آیند. [۱].

کاکانون و موزگووی یک ارزیابی سیستماتیک از هشت سیستم موجود آکادمیک و تجاری مرتبط با شناسایی سرقت ادبی را برای متون دانشجویی به انجام رساندند [۱] . سیستم‌های مورد ارزیابی در تحقیقات آن‌ها عبارت بودند از AntiPlagiarist (ACNP Software, ۲۰۱۰)، EVE۲ (Canexus، ۲۰۱۰)، Plagiarism-Finder (Mediaphor، ۲۰۱۰)، SafeAssignment (Sciworth، ۲۰۱۰)، SeeSources (۲۰۱۰)، Shrlock (Joy & Lock،

۱۹۹۹)، TurnitIn (iParadigms، ۲۰۱۰) و WCopyFind (BloomField، ۲۰۱۰). نتیجه اصلی تحقیقات آن‌ها آن بود که سیستم‌های کنونی شناساگر دارای نقایص زیادی هستند که می‌توان آن‌ها را به دو دسته عمده زیر تفکیک کرد:

- نقایص موجود در اجرای یک سیستم شناساگر بخصوص (مثلاً مسائل مرتبط با سهولت کاربری سیستم)؛
- مسائل ناشی از محدودیت‌های تکنولوژی‌های موجود برای شناسایی سرقت ادبی.

بنت به مطالعه فاکتورهای تحریک کننده دانشجویان برای سرقت ادبی پرداخت و دریافت که «ابزارها و فرصت‌ها» یکی از فاکتورهای موجود است [۳]. براساس این تحقیق این حقیقت که منابع به‌راحتی در اینترنت در دسترس هستند، امکان دستیابی سریع و ساده به حجم زیادی از اطلاعات از منابع مختلف را فراهم می‌آورد. همچنین، بسیاری از سایت‌های اینترنتی نیز وجود دارند که مقالات آماده را در اختیار دانشجویان قرار می‌دهند. سهولت دستیابی به مطالب از منابع آن‌لاین و استفاده از آن‌ها در تحقیقات آکادمیک، باعث افزایش شدید تعداد تحقیقات مرتبط با سرقت ادبی شده است (نیکسون و کسپرزاک ؛ نادلسون،اسکانلون و نیومن، [۱]).

۱-۳- سرقت ادبی چیست

«سرقت ادبی عبارتست از در اختیار گرفتن آثار فردی دیگر و انتقال آن‌ها به عنوان اثر شخصی. این جرم ارتباط نزدیکی را با جعل و دزدی ادبی دارد- عملکرد حالی که به معنای نقض قوانین کپی‌رایت می‌باشند»: دائرةالمعارف بریتانیکا.

سرقت ادبی در حقیقت یکی از جرایم الکترونیکی، مانند هک کامپیوتری، ویروس‌های کامپیوتری، اسپم، نقض قوانین کپی‌رایت و غیره، می‌باشد. سرقت ادبی را می‌توان در اختیار گرفتن یا تلاش برای در اختیار گرفتن یا استفاده (به صورت کلی یا جزئی) آثار فردی دیگر ، بدون اشاره و ارجاع به او به‌عنوان مالک اثر مطرح کرد. این کار می‌تواند به صورت یک copy و paste مستقیم، به واسطه اصلاح و تغییر برخی از لغات

متن اصلی که کتاب‌های اینترنتی، مجلات، روزنامه‌ها، تحقیقات، ژورنال ها، اطلاعات یا ایده‌های شخصی حاصل می‌شوند، باشد.

تعاریف بسیاری از موارد تشکیل‌دهنده سرقت ادبی وجود دارد که ما به برخی از آن‌ها نگاهی می‌اندازیم. با این حال بر اساس منابع تحقیقاتی ارائه‌شده در Plagiarism.org، مواردی که فوراً به ذهن می‌آیند را می‌توان به‌صورت زیر ارائه کرد:

- مطرح کردن اثر فردی دیگر به عنوان اثر شخصی
- کپی کردن کلمات یا ایده‌ها از نظر فردی دیگر بدون اشاره به منبع
- ارائه اطلاعات ناقص در مورد منبع یک نقل قول
- تغییر کلمات و در عین حال کپی کردن ساختار جمله یک منبع بدون اشاره به آن
- کپی کردن تعداد زیادی از کلمات یا ایده‌ها از منبع به طوری که اکثریت اثر شما را تشکیل دهد، خواه با ارجاع (اشاره به منبع) همراه باشد یا خیر [۴].

۱-۴- اهمیت سرقت ادبی

در برخی از سازمان‌های آکادمیک مانند دانشگاه‌ها، دانشکده‌ها و مؤسسات، تشخیص و جلوگیری از سرقت ادبی به یکی از چالش‌های آموزشی تبدیل شده است، زیرا اغلب دانشجویان و محققین در هنگام انجام پروژه‌ها و فعالیت‌های محول شده به آن‌ها، به این روش تقلب روی می‌آورند. این مسئله ناشی از دسترس پذیری منابع اینترنتی است. با دسترسی به این منابع، آن‌ها به راحتی می‌توانند با استفاده از یکی از موتورهای جستجو به دنبال موضوعات و تحقیقات گوناگونی بگردند بدون آن که نامی از مالک سند آورده شود. بنابراین برای همه حوزه‌های آکادمیک لازم است که از نرم‌افزارهای تشخیص سرقت ادبی در جهت جلوگیری یا حذف تقلب، کپی برداری و تغییرات سندهای دیگران توسط دانشجویان و افراد دیگر استفاده کنند.

برخی از انواع فعالیت‌های مرتبط با سرقت ادبی به راحتی می‌توانند با استفاده از نرم‌افزارهای تشخیص سرقت ادبی که اخیراً بوجود آمده‌اند و در بازار یا اینترنت موجود می‌باشند، کشف شوند. سرقت ادبی تنها مورد استفاده دانشجویان نیست بلکه برخی از محققین و اعضای کارمندان مؤسسات آکادمیک نیز از سرقت ادبی در مقالات منتشره خود به طور مستقیم یا غیرمستقیم استفاده می‌کنند.

برنامه‌های کامپیوتری زیادی برای تشخیص سرقت ادبی به کار گرفته می‌شود و ابزارهای شناساگر گوناگونی توسط محققین به‌وجود آمده‌اند، اما آن‌ها همچنان دارای محدودیت‌هایی هستند و نمی‌توانند به درستی تشخیص دهنده و اثبات کننده سرقت ادبی باشند؛ بلکه صرفاً شباهت‌ها را نشان می‌دهند . با این حال بسیاری از دانشگاه‌ها و مراکز تحقیقاتی هنوز هیچ اقدامی را برای مقابله با سرقت ادبی انجام نداده‌اند و این مسئله باعث رشد بیشتر سرقت ادبی شده است. علیرغم این موارد، حتی با استفاده از نرم‌افزارهای جدید شناساگر نیز نمی‌توان سرقت ادبی راه ۱۰۰٪ محو کرد.

کپی رایت و جنبه‌های حقوقی استفاده از اسناد منتشر شده نیز می‌توانند تحت پوشش نرم‌افزارهای شناسایی سرقت ادبی قرار گیرند تا مشخص شود که آیا یک فرد از اسناد دیگران به صورت قانونی استفاده کرده است یا غیر قانونی و اینکه آیا فرد مجوز لازم را از طرف مالک برای استفاده از این سند داشته است یا خیر.

شناسایی سرقت ادبی همچنین یکی از مهم‌ترین مسائل برای سازمان‌ها، مراکز تحقیقاتی و کنفرانس‌ها است؛ آن‌ها از ابزارهای پیشرفته شناسایی سرقت ادبی برای اطمینان از این موضوع استفاده می‌کنند که اسناد آن‌ها مورد سرقت ادبی قرار نگیرند، و در نتیجه حق کپی رایت ناشران محفوظ بماند.

یک تحقیق (منتشر شده در ژوئن ۲۰۰۵) که به صورت بخشی از پروژه ارزیابی مرکز یکپارچگی آکادمیک انجام شد، نشان داد که ۴۰٪ از دانشجویان به انجام فعالیت‌های مرتبط با سرقت ادبی اقرار کرده‌اند ؛ در حالی که این مسئله در سال ۱۹۹۹، ۱۰٪ بود [۴] . یک مطالعه جمعی دیگر که توسط استاد دانشگاه Rutgers در سال ۲۰۰۳ انجام شد، گزارش داد که ۳۸٪ از دانشجویان در سرقت ادبی آن‌لاین مشارکت داشتند [روتگار، ۲۰۰۳]. این ارقام هشداردهنده بیانگر افزایش تدریجی سرقت ادبی می‌باشد. نسل جدید

نسبت به گذشته از تکنولوژی آگاه‌تر هستند. سرقت ادبی هم‌اکنون محدود به یک copy و paste کردن صرف نیست؛ تکنولوژی‌های مترادف سازی و ترجمه در حال ایجاد ابعاد جدیدی در سرقت ادبی هستند.

هم‌اکنون سرقت ادبی مهم‌ترین سوءرفتار آکادمیک محسوب می‌شود؛ آکادمی ها و مؤسسات سراسر جهان در حال شروع فعالیت‌هایی برای آموزش به دانشجویان و مدرسین در جهت توصیف و تحلیل انواع سرقت ادبی و چگونگی اجتناب از آن می‌باشند.

نادلسون تحقیقی را در میان ۷۲ فرد آکادمیک در خصوص مسائل مرتبط با سوءرفتار آکادمیک انجام داد و گزارش کرد که ۵۷۰ نمونه از سرقت ادبی مظنون مشاهده شده است [۱]. اغلب موارد گزارش شده شامل «سرقت ادبی تصادفی/غیرعمدی» بوده‌اند که در این میان ۱۳۴ مورد مربوط به دانشجویان دورهٔ لیسانس و ۳۹ مورد مربوط به دانشجویان فارغ‌التحصیل بود. بعلاوه، افراد آکادمیک گزارش کردند که تعداد زیادی از دانشجویانی را سراغ دارند که مقالات کپی شده از اینترنت را ارائه کرده‌اند. همچنین مواردی از «سرقت ادبی هدفمند»، « تقلب های آزمون کلاسی» و « تقلب های آزمون‌های خانگی» نیز گزارش شدند.

بعلاوه سرقت ادبی مشکل بزرگی در دوره‌ای برنامه‌نویسی است. کالوین تحقیقی را از سرقت ادبی کد منبع انجام داد و در آن داده‌هایی را از ۵۵ دانشکده کامپیوتر در بریتانیا به دست آورد[۵]. او دریافت که ۵۰٪ از ۲۹۳ فرد آکادمیک مورد بررسی عقیده داشتند که سرقت ادبی در سال‌های اخیر افزایش یافته است. بعلاوه، ۲۲ نفر از ۴۹ پاسخ دهنده برآوردهایی از ۲۰٪ تا ۵۰٪ را در خصوص دانشجویان مشارکت کننده در فعالیت‌های سرقت ادبی در دوره‌های ابتدایی برنامه‌نویسی در ذهن داشتند.

۱-۵- مقیاس مشکل

یک بررسی اولیه در میان ۳۸۰ دانشجوی دوره کارشناسی که توسط هاینز، دیکهوف، لابف و کلارک در سال ۱۹۸۶ انجام شد نشان داد که هر چند حدود نیمی از دانشجویان به انواع مختلف تقلب می‌کنند، اما تنها ۱٫۳٪ از این موارد، تشخیص داده می‌شود [۶]. بررسی‌های اخیر نشان داده‌اند که در دو دههٔ گذشته، سرقت

ادبی در میان دانشجویان به دلیل سهولت دسترسی و اشتراک گذاری پاسخ‌های تکالیف و نیز مطالب تحقیقاتی، بسیار رایج شده است.

سرقت ادبی در تکالیف برنامه‌نویسی یک مشکل روزافزون در محیط‌های آکادمیک است. کد منبع را می‌توان به طرق مختلف مانند اینترنت، بانک‌های کد منبع و کتاب‌های راهنما به دست آورد. در خصوص موارد مظنون به سوءرفتار آکادمیک، تحقیقات مشخص گردند که به دلیل مشکلات روش‌های رسمی دانشگاه‌ها، افراد آکادمیک ترجیح می‌دهند که مسائل دخیل در سرقت ادبی را به صورت غیررسمی رفع و رجوع کنند. این مشکلات اساساً ناشی از دو مسئله است-افراد آکادمیک احساس می‌کنند که شواهد کافی را برای گزارش این سرقت ها ندارند، و اینکه پیگیری رسمی موارد سوءرفتار آکادمیک می‌تواند وجهه آن‌ها را به عنوان افراد آکادمیک به صورت منفی تحت تأثیر قرار دهد. بعلاوه، کیت-اشپیگل دریافته‌اند که افراد آکادمیک عموماً سیاست‌های دانشگاه را در مورد سرقت ادبی دنبال نمی‌کنند زیرا از مواجهه با دانشجویان و عدم پشتیبانی شیوه‌های دانشگاهی بیم دارند. آن‌ها همچنین دریافته‌اند که افراد آکادمیک با توجه به تأثیراتی که ممکن است روش‌های رسمی پیگیری سرقت ادبی بر دانشجویان داشته باشد، با آن‌ها با ملایمت بیشتری رفتار می‌کنند.

۱-۶- دسته‌بندی‌های سرقت ادبی

دسته‌های گسترده‌تر سرقت ادبی شامل موارد زیر می‌باشند:

- تصادفی: به دلیل عدم اطلاع از چگونگی سرقت ادبی و عدم درک چگونگی استفاده از روش‌های ارجاع (اشاره به منبع) مورد استفاده در یک سازمان
- غیرعمدی: گستردگی اطلاعات موجود می‌تواند بر ایجاد شباهت‌هایی میان تفکر و ایده‌های به دست آمده از مطالب مکتوب یا شفاهی افراد اثرگذار باشد
- عمدی: عملکرد تعمدی کپی برداری کلی یا جزئی از آثار فردی دیگر بدون اشاره به منبع اثر

- سرقت ادبی شخصی: استفاده از آثار منتشرشده شخصی به شکلی دیگر بدون ارجاع به اثر اصلی [Wikipedia:Plagiarism ۲۰۰۶].

به طور دقیق‌تر، و از جنبه روش کار، سرقت ادبی می‌تواند اشکال مختلفی به خود بگیرد ؛ که شامل موارد زیر می‌باشد.

- کپی برداری کلمه به کلمه، که شامل کپی برداری مستقیم جملات یا مجموعه‌ای از جملات از آثار دیگران بدون ارجاع به منبع اصلی می‌باشد.
- تغییر الفاظ، که شامل بازنویسی نزدیک (تنها تغییر برخی از کلمات بدون تغییرات کلی) متن مکتوب بدون ارجاع مناسب به منبع اثر می‌باشد.
- سرقت ادبی از منابع ثانویه، که شامل ارجاع و اشاره به منبع ثانویه یک متن بدون جستجو به دنبال منبع اصلی می‌باشد.
- سرقت ادبی شکل یک منبع زمانی اتفاق می‌افتد که ساختار یک آرگومان در یک منبع بدون استفاده سیستماتیک از ارجاعها از یک منبع ثانویه انجام می‌شود. این گونه سوء رفتار شامل جستجوی مرجع ها و دنبال کردن ساختاری مشابه منبع ثانویه می‌باشد.
- سرقت ادبی ایده‌ها، که شامل استفاده از ایده‌های مطرح‌شده در یک متن منبع بدون هیچ گونه اشاره مستقیم به کلمات یا شکل منبع می‌باشد.
- سرقت ادبی مشهود یا مالکیتی، یعنی در اختیار گرفتن اثر دیگران و معرفی آن بعنوان اثر خود.

۱-۷- تحقیقات انجام شده از نظر دسته بندی روش‌های شناسایی

۱-۷-۱- اثر انگشت ها[۱]

بر مبنای تحقیقات الظهرانی پیرامون روش‌های تشخیص سرقت ادبی، روش‌های رایج مبتنی هستند بر کاراکترها برای مقایسه اسناد مذنون با اسناد اصلی. رشته همسان می‌تواند به طور دقیق یا جزئی با استفاده از روش انطباقی کاراکتر تشخیص داده شود[۷]. لیون مقایسه متنی را بر مبنای n-گرم های کلمات انجام می‌دهد ؛ و با ارجاع به آن‌ها، متن مورد نظر به دو مجموعه از تری-گرم ها تقسیم می‌شود تا مورد مقایسه قرار گیرد. میزان تری-گرم های مشترک به منظور بررسی موارد بالقوه سرقت ادبی، شناسایی می‌شود[۷]. کنگ جمله را به عنوان واحد مقایسه در نظر می‌گیرد تا شباهت های محلی میان آن‌ها را بررسی کند. او تمایزی را میان کپی برداری دقیق جملات، درون گذاری کلمه، حذف کلمات و تعویض آن‌ها با کلمات هم معنی قایل می‌شود[۷]. هاینتزه، برودرو کروسزتی یک روش اثر انگشت را برای یافتن انطباق رشته‌ای و تشخیص سرقت ادبی بر مبنای نسبت به اثر انگشت های رایج پیشنهاد کردند. این روش‌ها نتایج خوبی را به دست می‌دهند، اما نمی‌توانند بخش‌های حاوی کلمات تعویض شده با کلمات هم معنی یا متفاوت را تشخیص دهند[۷].

۱-۷-۲- شباهت رشته‌ای[۲]

برین یک سیستم تشخیص سرقت ادبی را با نام COPS (سیستم محافظت در برابر کپی) از پروژه کتابخانه دیجیتال استنفرد ایجاد کرد، که می‌تواند تداخل اسناد را بر مبنای جملات و انطباق رشته‌ای تشخیص دهد[۷]. نقص اصلی این روش، ناتوانی در بررسی کلمات منفرد و در نظر گرفتن کل جمله به عنوان یک بخش منفرد است. نقایص روش COPS توسط شیواکومار و گارسیا- مولینا برطرف شد و روش جدید به دست آمد با نام روش آنالیز کپی استانفورد (SCAM) تا بتواند COPS را با استفاده از مدل تناوب نسبی (RFM)

---

[۱] Fingerprints

[۲] String Similarity

برای نشان دادن زیرمجموعه‌های کپی برداری شده، بهبود بخشد[۷]. RFM عبارتست از یک معیار شباهت نامتقارن برای تشخیص سرقت ادبی. مزیت اصلی SCAM آن است که می‌تواند شباهت‌های متداخل را میان اجزای جمله ها تشخیص دهد ، اما عبارات بسیاری وجود دارند که می‌توانند در مقایسه تقسیمات میان اسناد، گمراه‌کننده باشند. روش پیشنهادی توسط هانت و سزیمانسکی و اقتباس شده توسط چو و سلیم ، بلندترین زیر-دنباله مشترک (LCS) نام دارد. روش LCS یکی از شیوه‌های مورد استفاده در ROUGE است که در حقیقت یک روش ارزیابی خلاصه معروف می‌باشد[۷]. با داشتن دو دنباله X و Y، بلندترین زیر-دنبالهٔ مشترک برای آن‌ها می‌تواند زیر دنباله مشترک دارای بلندترین ماکزیمم باشد. وایت و جوی یک الگوریتم جدید را برای تشخیص سرقت ادبی پیشنهاد کردند که می‌تواند تداخل های پیچیده (مانند تغییرات کلمات، بازآرایی، یکپارچه سازی، و تفکیک جملات) و نیز کپی برداری مستقیم را تشخیص دهد.

۱-۷-۳- روش‌های مبتنی بر ساختار[۳]

توجه به این نکته اهمیت دارد که همه مطالعات پیش‌گفته، روش‌های مبتنی بر کاراکتر را توصیف کردند. در حقیقت همه این روش‌ها روی ویژگی‌های کلمات متن در یک سند متمرکز بودند. با این حال بسیاری از مطالعات، روش‌های مختلفی را در حوزه تشخیص سرقت ادبی بر مبنای ویژگی‌های ساختاری متن موجود در سند، مانند سربرگ، تعداد بخش‌ها، پاراگراف ها و منابع پیشنهاد نموده‌اند. ویژگی‌های ساختار درختی یکی از مطالعات اخیر با تمرکز بر ویژگی‌های ساختاری است. این یک مدل از نقشه‌های خود-آرای چند لایه (ML-SOM) برای اسناد متنی می‌باشد. چو و رحمان از یک نمایش ساختار درختی با ML-SOM برای تشخیص سرقت ادبی و بازیابی اطلاعات استفاده کردند. ایده آن‌ها مبتنی بر دو مرحله بود، لایه بالا ولایه پایین. لایه بالا نشان‌دهنده دسته‌بندی و بازیابی اسناد است در حالی که لایه پایین از یک ضریب شباهت کسینوسی برای تعیین شباهت‌های میان متون درنتیجه تشخیص سرقت ادبی استفاده می‌کند[۷].

---

[۳] Structure-Based Methods

۱-۷-۴- روش‌های خوشه بندی[۴]

روش خوشه بندی سند یکی از روش‌های بازیابی اطلاعات است که در بسیاری از حوزه‌ها مانند خلاصه سازی متن، دسته‌بندی متن، و تشخیص سرقت ادبی مورد استفاده قرار می‌گیرد. از آن برای بهسازی داده‌های بازیابی شده با استفاده از کاهش زمان جستجو در موقعیت سند برای خلاصه سازی متن و کاهش زمان مقایسه در تشخیص سرقت ادبی استفاده می‌شود. شیوه دیگری که توسط سی و همکاران، و زینی پیشنهاد شده، از کلمات خاص (کلمات کلیدی) برای یافتن خوشه های مشابه میان اسناد استفاده می‌کند. زینی نیز اشاره کرد که روش‌های اثر انگشت اساساً مبتنی بر استفاده از k-گرم ها است[۷] . از آنجا که فرایند اثر انگشتی تقسیم کننده اسناد به گرم های (grams) دارای طول k هستند، اما اثر انگشتی های سند را می‌توان به منظور تشخیص سرقت ادبی، مقایسه کرد. روش‌های مبتنی بر اثر انگشتی برای خوشه بندی اسناد برای خلاصه سازی مجموعه‌ای از اسناد و ایجاد یک مدل اثر انگشتی برای آن‌ها مورد استفاده قرار گرفتند.

۱-۷-۵- مدلهای نحوی[۵]

الهادی و التوبی یک روش تشخیص کپی برداری را برای ساختارهای نحوی اسناد پیشنهاد کردند. این روش نگاهی دارد به استفاده از برچسب‌های اقسام نحوی جمله (PoS) برای نمایش ساختار متن به صورت مبنایی برای مقایسه و آنالیز بیشتر[۷]. این روش می‌تواند اسناد را با استفاده از برچسب‌های PoS مرتب و رتبه‌بندی کند. الهادی و التوبی روش خود را برای محاسبه شباهت میان اسناد و رتبه بندی آن‌ها بر مبنای مرتبط ترین اسناد استخراج شده ارتقا دادند.

برخی از مطالعات مانند تحقیق کروتشف و سربیان ترکیبات نحوی دو متن را بر مبنای الگوریتم فاصله نرمال سازی شده Lempel-Ziv (LZ) متراکم سازی کردند و شباهت‌های میان اطلاعات توپولوژیکی مشترک ارائه

---

[۴] Cluster Based Methods

[۵] Syntactic Models

شده توسط متراکم ساز را محاسبه نمود. این روش می‌توانست شباهت‌های میان اسناد متنی را حتی در صورتی تشخیص دهد که آن‌ها دارای الفاظ متفاوت می‌باشند[۷]. بعلاوه، این روش و سایر روش‌هایی مانند چیزی که توسط االهادی و همکارانش پیشنهاد شد، مبتنی بودند بر کاهش متن با استفاده از نشانه سازی و حذف کلمات تکراری و تنها مجموعه‌های کوچک‌تری از برچسب‌های نحوی را مدنظر قرار می‌دهند.

۱-۷-۶- روش‌های چند زبانی[۶]

یک روش چند زبانی برای تشخیص اسناد مذنون توسط گروزیا و پوپسکو پیشنهاد شد. شباهت میان اسناد اصلی و مذنون با استفاده از یک مدل آماری ارزیابی میشود. این مدل می‌تواند این احتمال را تعیین کند که یک سند مذنون با سند اصلی مرتبط است یا خیر، بدون توجه به ترتیب عبارات در دو سند. روش آن‌ها با یک مجموعه داده متنی در انگلیسی و اسپانیایی ترکیب شده با شباهت‌های میان زبانی را تشخیص دهد. این روش نیازمند ایجاد یک مجموعه داده میان زبانی است که همان‌طور که گروزیا و پوپسکو گفته‌اند، کار بسیار دشواری است.

۱-۷-۷- روش‌های معنایی (صرفی)[۷]

بسیاری از محققین کارهای زیادی را برای محاسبه شباهت‌های معنای میان کلمات و ارتباط میان اسناد با استفاده از WordNet انجام داده‌اند. معیارهای مبتنی بر اطلاعات می‌توانند شباهت معنایی میان دو کلمه را با محاسبه درجه ارتباط میان آن‌ها با استفاده از اطلاعات یک دیکشنری یا فرهنگ لغت تشخیص دهند. در اینجا از روابط و سلسله‌مراتب یک فرهنگ لغت استفاده می‌شود که عموماً یک پایگاه داده لغوی ایجاد شده به صورت دستی مانند WordNet می‌باشد. تاآنجاکه ما می‌دانیم، اسناد متنی می‌توانند به صورت یک گراف نمایش داده شوند. سرقت ادبی می‌تواند به واسطه انطباق گراف میان مفاهیم کلمات نشان داده شود. یک روش معنای تشخیص سرقت ادبی که توسط الظهرانی و سلیم پیشنهاد شد، از شباهت رشته‌ای مبتنی بر

---

[۶] Cross-Language Methods

[۷] Semantic Methods

معنای فازی استفاده می‌کند[۷]. این روش به واسطه چهار مرحله اصلی انجام گرفت. مرحله اول عبارتست از پیش پردازش که شامل نشانه سازی، حذف کلمات مشابه و ریشه‌یابی. مرحله دوم عبارت است از بازیابی فهرستی از اسناد منتخب برای هر سند مذنون با استفاده از ضریب Jaccard و الگوریتم Shingling. سپس اسناد مذنون به صورت جمله به جمله با اسناد منتخب مقایسه می‌شوند. این مرحله دربرگیرنده محاسبهٔ درجهٔ شباهت فازی است که دامنه‌ای از ۰ برای دو جمله کاملاً متفاوت را ۱ برای دو جمله کاملاً یکسان دارد. دو جمله در صورتی علامت‌گذاری می‌شوند که دارای یک نمره شباهت فازی در بالای یک آستانه معین باشند. مرحله آخر، پردازش واپسین است که در آن جملات متعاقب به یکدیگر پیوند داده می‌شوند تا یک پاراگراف یا بخش منفرد را شکل دهند.

روش متفاوتی از تشخیص سرقت ادبی به صورت معنایی توسط چو و سلیم پیشنهاد شد [۷]. روش پیشنهادی می‌تواند شباهت میان اسناد مذنون و اصلی را بر مبنای مستندات جمله محاسبه کند. هر یک از این مستندات با استفاده از یک درخت تجزیه گر استانفورد (SPT) استخراج می‌شود. درجهٔ شباهت میان مستندات استخراج شده با استفاده از فرهنگ لغت WordNet محاسبه می‌شود. نقص این روش آن است که نمی‌تواند همه بخش‌های جملات را تشخیص دهد بلکه تنها فاعل، فعل و مفعول را بررسی می‌کند.

۱-۸- خلاصه فصل اول

سرقت ادبی عبارتست از در اختیار گرفتن آثار فردی دیگر و انتقال آن‌ها به عنوان اثر شخصی. در برخی از سازمان‌های آکادمیک مانند دانشگاه‌ها، دانشکده‌ها و مؤسسات، تشخیص و جلوگیری از سرقت ادبی به یکی از چالش‌های آموزشی تبدیل شده است، زیرا اغلب دانشجویان و محققین در هنگام انجام پروژه‌ها و فعالیت‌های محول شده به آن‌ها، به این روش تقلب روی می‌آورند این مسئله ناشی از دسترس پذیری منابع اینترنتی است. برنامه‌های کامپیوتری زیادی برای تشخیص سرقت ادبی به کار گرفته می‌شود و ابزارهای شناساگر گوناگونی توسط محققین به‌وجود آمده‌اند، اما آن‌ها همچنان دارای محدودیت‌هایی هستند و نمی‌توانند به درستی تشخیص دهنده و اثبات کننده سرقت ادبی باشند؛ بلکه صرفاً شباهت‌ها را نشان

می‌دهند . بررسی‌های اخیر نشان داده‌اند که در دو دهۀ گذشته، سرقت ادبی در میان دانشجویان به دلیل سهولت دسترسی و اشتراک گذاری پاسخ‌های تکالیف و نیز مطالب تحقیقاتی، بسیار رایج شده است. تحقیقات انجام شده از نظر دسته بندی روش‌های شناسایی عبارتتند از :

روش های اثر انگشت : این روش نتایج خوبی را به دست می‌دهند، اما نمی‌توانند بخش‌های حاوی کلمات تعویض شده با کلمات هم معنی یا متفاوت را تشخیص دهند.

روش های خوشه بندی : روش‌های مبتنی بر اثر انگشتی برای خوشه بندی اسناد برای خلاصه سازی مجموعه‌ای از اسناد و ایجاد یک مدل اثر انگشتی برای آن‌ها مورد استفاده قرار گرفتند.

مدل های نحوی : این روش می‌توانست شباهت‌های میان اسناد متنی را حتی در صورتی تشخیص دهد که آن‌ها دارای الفاظ متفاوت می‌باشند. بعلاوه، کاهش متن با استفاده از نشانه سازی و حذف کلمات تکراری و تنها مجموعه‌های کوچک‌تری از برچسب‌های نحوی را مدنظر قرار می‌دهند.

روش‌های چند زبانی : این روش شباهت‌های میان زبانی را تشخیص میدهد. این روش نیازمند ایجاد یک مجموعه داده میان زبانی است که کار بسیار دشواری است.

روش‌های معنایی (صرفی) : شباهت معنایی میان دو کلمه را با محاسبه درجه ارتباط میان آن‌ها با استفاده از اطلاعات یک دیکشنری یا فرهنگ لغت تشخیص دهند. در اینجا از روابط و سلسله‌مراتب یک فرهنگ لغت استفاده می‌شود که عموماً یک پایگاه داده لغوی ایجاد شده به صورت دستی مانند WordNet می‌باشد.

فصل دوم : سرقت ادبی در زبان ها

۲-۱- سرقت ادبی درون زبانی

سرقت ادبی به نام زبانی در حقیقت چیزی است که آن را به صورت دریافت عبارات، مفاهیم و جملات از متن دیگر و انتقال آن به متن خود در حیطه یک زبان منفرد تعریف می‌کنند. روش‌های بسیاری برای تشخیص سرقت ادبی توسط محققین در سال‌های گذشته به وجود آمده‌اند شما در اینجا به برخی از آن‌ها اشاره می‌کنیم که مرتبط با تشخیص ادبی در حیطه زبان منفرد هستند.

ابتدا، برخی از محققین از روش‌های تشخیص کپی متن «زبان طبیعی» استفاده کردند که این روش در دهه ۹۰ پدیدار شد و در آن سه روش تشخیص سرقت ادبی مطرح شد [۸].

۲-۱-۱ روش‌های مبتنی بر گرامر[۸]

روش مبتنی بر گرامر[۹] یکی از مهم‌ترین روش‌های مورد استفاده برای تشخیص سرقت ادبی است که بر ساختار گرامری اسناد تمرکز می‌کند و از یک شیوه انطباق رشته‌ای برای شناسایی و اندازه‌گیری شباهت میان اسناد بهره می‌گیرد. روش‌های مبتنی بر گرامر برای تشخیص کپی‌های دقیق بدون هرگونه تغییرات ، مورد استفاده قرار می‌گیرند اما نمی‌توانند متون کپی شده و در عین حال اصلاح شده را تشخیص دهند. این یکی از محدودیت‌های این روش‌ها است. در این زمینه می‌توان به روش‌های زیر اشاره کرد:

- هوانگ یک روش تشخیص برای صفحات وب را بر مبنای الگوریتم LCS طریق یافتن بزرگ‌ترین رشته مشترک میان دو صفحه برای محاسبهٔ شباهت میان آن‌ها ارائه کرد[۹].
- اشلایمر ، ویلکرسون و آیکن از روش kگرم تداخلی برای دریافت hashهای اسناد و به دست آوردن اثر انگشتها با کاهش تعداد hashها در سند استفاده کردند و سپس از اثر انگشت برای هر سند به دست آمده بهره گرفتند و نرخ‌های شباهت میان این اسناد را شمردند[۹].

---

[۸] Grammer-based methods

- breaking Hash و DCT نیز از روش‌های گرامر-محور هستند؛ تنها تفاوت میان آن‌ها چگونگی دریافت اثر انگشت‌های میان اسناد است.

۲-۱-۲- روش معنایی[۹]

روش معنایی نیز یکی از روش‌های مهم برای تشخیص سرقت ادبی است و مبتنی بر شباهت‌های میان اسناد به واسطه مدل فضای برداری می‌باشد. این روش همچنین می‌تواند حشویات کلمات را در سند شمرده و محاسبه کند و سپس از اثر انگشتهای اسناد دیگر برای یافتن شباهت استفاده کند.

۲-۱-۳- روش ترکیبی گرامر-معنایی

روش ترکیبی گرامر-معنایی یکی از مهم‌ترین روش‌ها در تشخیص سرقت ادبی برای زبان‌های طبیعی است. این روش، که در دستیابی به نتایج بهتر و مناسب‌تر تأثیرگذاری خود را نشان داده است، برای متون کپی شده، تغییر یافته، و بازنویسی شده ای مناسب است که نمی‌توان آن‌ها را از طریق روش های گرامر-محور و نیز روش‌های معنایی بررسی کرد.

دوم آنکه برخی از روش‌ها از ساختارهای شاخص-محور بخصوصی استفاده کردند:

- ملکولم و لین از سیستم تشخیص سرقت Ferret بهره گرفتند که مبتنی است بر تری-گرم های کلمات مشترک. [۹]
- باسیل متن ها را به صورت یک دنباله طول کلمات رمز گذاری کرد و از شاخص فاصله n-گرم مبتنی بر بردار جریان نزولی برای انتخاب بخش معیار (کاندیدا) استفاده کرد. [۹]
- کاسپرزاک شینگل های متنی مشترک را در فرایند پیش-گزینش مورد استفاده قرار داد و اشکربینین و بوتاکف [۹] از اثر انگشت های مبتنی بر hash برای بازیابی پخش گزینشی استفاده نمودند. [۹]

---

[۹] Semantic methods

سوم، روش تشخیص سرقت ادبی خارجی :

روش خارجی تشخیص سرقت ادبی مبتنی است بر یک مجموعه داده مرجع متشکل از اسنادی که در آن‌ها متون ممکن است سرقت شده باشند. یک متن می‌تواند از پاراگراف ها ، یک بلوک از کلمات با اندازهٔ ثابت، یک بلوک از جملات و غیره تشکیل شود. یک سند مذنون به واسطه جستجو برای متونی کپی برداری شده که در مجموعه داده مرجع قرار دارند، از نظر سرقت ادبی مورد بررسی قرار می‌گیرند. سپس یک سیستم خارجی سرقت ادبی این یافته‌ها را برای یک کنترلگر انسانی گزارش می‌کند و او تشخیص می‌دهد که آیا متون شناسایی شده مورد سرقت قرار گرفته‌اند یا خیر. یک راهکار برای این مسئله می‌تواند مقایسه هر متن در سند مذنون با هر متنی از یک سند در یک مجموعه داده مرجع باشد.

مزایا و معایب : مجموعه داده مرجع باید به حد کافی بزرگ باشد تا حداکثر میزان ممکن از متون سرقت شده یافت شوند. این موضوع باعث افزایش شدید زمان اجرا میشود.

نمونه‌هایی از تحقیقات انجام شده برای تشخیص سرقت ادبی خارجی عبارتند از:

- تشخیص اتوماتیک جهت‌گیری سرقت ادبی؛ از این روش برای تعیین جهت سرقت ادبی، استفاده شد. گروزیا و پوپسکو از تعمیمی از روش Encoplot برای این منظور بهره گرفتند و آن را روی مجموعه بزرگی از سرقت‌های ادبی مصنوعی آزمودند و نشان دادند که در بزرگ‌ترین مجموعه داده ای سرقت ادبی موجود تا به امروز، مشکل جهت‌گیری سرقت ادبی با دقت نسبتاً بالا (در حدود ۷۵٪) برطرف شد، اما آن‌ها این روش را در مورد زبان طبیعی نیازمودند [۹].
- روش تشخیص اتوماتیک سرقت ادبی به شکل خارجی از شباهت‌های متنی استفاده کرد، این روش در مرحلهٔ پیش پردازش به کار گرفته شد. وانیا و آدریانی اسناد بازیابی شده را به متون S تقسیم کردن و هر متن دارای ۲۰ جمله بود و سپس سرقت ادبی را با شناسایی تعداد لغات متداخل میان متون مذنون و مرجع شناسایی کردند. [۹]

- دوی ، رائو ، رام و آکیلاندسواری یک الگوریتم را برای شناسایی سرقت ادبی خارجی در رقابت PAN-۱۰ مطرح کردند. [۹] این الگوریتم دارای دو مرحله بود: مرحله اول شناسایی اسناد مشابه و بخش سرقت شده برای سند مذنون با مدل فضای برداری (VSM) و شاخص شباهت کسینوسی، و مرحله دوم شناسایی نواحی سرقت شده در متن مذنون با استفاده از نسبت Chunk. اما پیش پردازش اسناد انجام نشد.
- مارکوس موهر، کرن ، زکنر و گرانیتزر سیستم ترکیبی خود را برای رقابت PAN در ۲۰۱۰ CLEF مطرح نمودند. [۹] سیستم آن‌ها تشخیص سرقت ادبی را برای متون ترجمه شده و ترجمه نشده و نیز متون اسناد سرقت شده داخلی انجام می‌داد؛ روش تشخیص سرقت ادبی خارجی به صورت یک مسئله بازیابی اطلاعات فرمول بندی شد، و از فرا-پردازش شهودی برای دستیابی به نتایج شناسایی نهایی استفاده کرد.
- زکنر و همکاران ، یک مدل استاندارد از IR متنی را برای گزینش بخش کاندیدا به کار گرفتند. اسناد مرجع در سطح جمله، شاخص گذاری شدند و جملات اسناد مذنون به عنوان عبارات جستجو به کار گرفته شدند. شباهت از طریق شاخص کسینوس برآورد شد. [۹]

۲-۱-۴- تشخیص سرقت ادبی بر مبنای اطلاعات ساختاری

در این روش فرض می‌شود که D_x مجموعه‌ای از اسناد مذنون و D_s مجموعه‌ای از اسناد مرجع باشند. اولین کار عبارت است از تعیین اینکه آیا یک سند مذنون مورد سرقت قرار گرفته است یا خیر. باید همه منابع سرقت، مانند اسناد منبع (زیر مجموعه‌ای از D_s) و مرزهای دقیق متون سرقت شده در اسناد مذنون و مرجع شناسایی شود. بعلاوه، بهتر است نمره‌ای به هر متن سرقت شده شناسایی شده اختصاص داده شود که نشان دهنده درجه سرقت ادبی در آن است. این نمره را می‌توان برای مرتب کردن متن‌های شناسایی شده از «کپی‌های دقیق» تا «متن‌های نسبتاً مرتبط»به کار گرفت. [۱۱]

نمایش متن

نمایش متون براساس روش پیشنهادی، مبتنی است بر n-گرم های کلمات کلیدی (SWNG). با داشتن یک سند و فهرستی از کلمات کلیدی، متن به نمایشی از این کلمات کلیدی در سند تقلیل پیدا می‌کند. همه نشانه‌های دیگر حذف می‌شوند.

در تصویر ۱، نمونه‌ای از نمایش SWNG را مشاهده می‌کنید. متن سمت راست، متن اصلی بوده و متن سمت چپ ، متنی است که مظنون به سرقت ادبی می‌باشد.

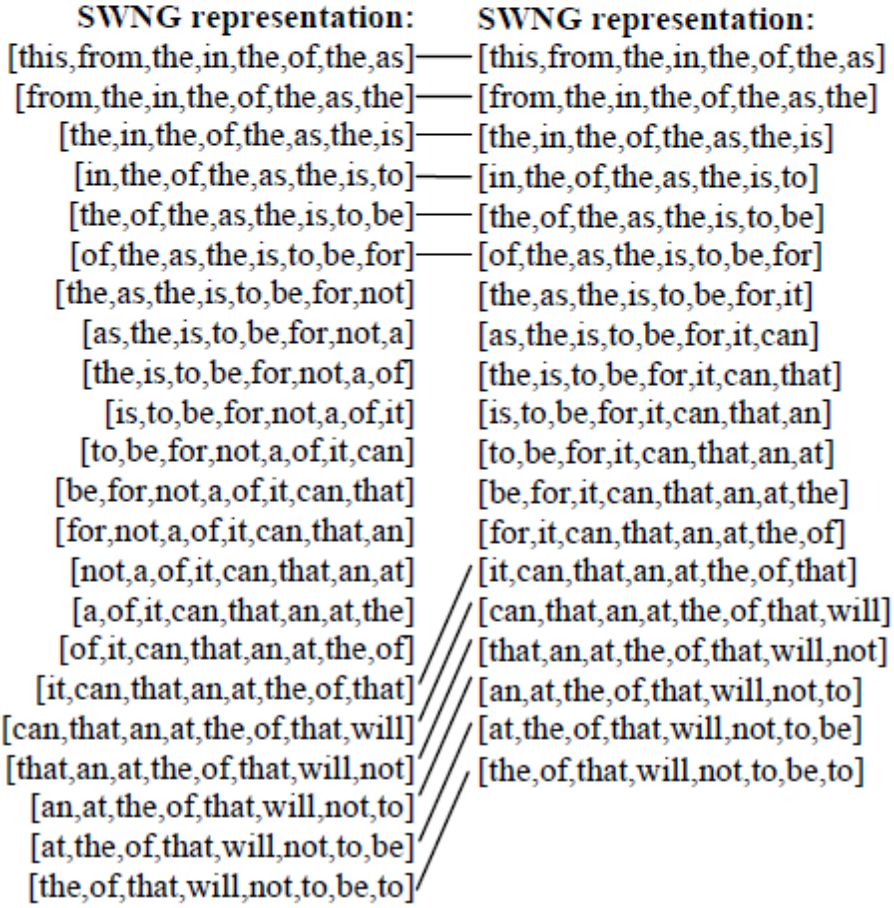
مزایا و معایب: این روش میتواند بازدهی بسیار بالایی را در مورد سرقت های ادبی دشوار بدست آورد که در آنها تغییرات قابل توجهی در متن رخ داده است و کلمات و عبارات با مترادف های آنها جایگزین شده اند. کارکرد این روش بسیار آسان است و نیازمند حداقل منابع و حداقل هزینه پردازش متن می باشد.

Suspicious passage:

This came into existence likely from the deviance in the time-period of the particular billet. As the premier is to be nominated for not more than a period of four years, it can infrequently happen that an ample wage, fixed at the embarkation of that period, will not endure to be such to its end.

Original passage:

This probably arose from the difference in the duration of the respective offices. As the President is to be elected for no more than four years, it can rarely happen that an adequate salary, fixed at the commencement of that period, will not continue to be such to its end.



تصویر ۱: نمونه‌ای از یک مورد سرقت ادبی و نمایش n-گرم کلمات کلیدی (۸- گرم مشترک متون با خط‌ها نشان داده شده‌اند) [۱۱]

بازیابی متن گزینشی (کاندیدا)[۱۰]

اولین مرحله مهم در تشخیص سرقت ادبی با این روش، بازیابی یک زیرمجموعه از D_s که متشکل از موارد احتمالی سرقت ادبی در یک سند مظنون باشد. این روش شامل مقایسه گسترده سند مظنون با هر تعداد از D_sها برای شناسایی شباهت‌های محلی می‌باشد. معمولاً تعداد سندهای منبع برای یک سند مظنون معین از قبل مشخص نیست.

تشخیص مرز متنی

هنگامی که مجموعه‌ای از اسناد منبع که با یک سند مظنون منطبق هستند، تعیین می‌شوند، مرحله بعدی عبارتست از انجام یک آنالیز دقیق برای ارزیابی مرزهای[۱۱] دقیق متن‌های سرقت شده در هر دو سند منبع و مظنون. فرض کنیم که $D_{rx} \subseteq D_s$ نشاندهنده مجموعه اسناد منبع باشد که برای سند مظنون d_x بازیابی شده‌اند. هدف این روش یافتن SWNGهایی در پروفایل های d_x و هر $d_s \in D_{rx}$ و ایجاد جملات ماکسیمال از آن‌ها است که با متن متناظر هستند.

در تصویر ۲، الگوریتم مربوط به تشخیص مرزهای متنی نشان داده شده است.

---

[۱۰] Candidate text retrieval

[۱۱] Boundaries

```
Input: dx, a suspicious document
       ds, a source document
       n2, length of stopword sequences
       θg, threshold of maximum gap allowed
Output: a set of detections

detectPassageBoundaries(dx,ds,n2,θg)
1. Px=profile(n2,dx);
2. Ps=profile(n2,ds);
3. [M1,M2]=match(Px,Ps);
4. InitPlagPass=findPassages(M1,θg);
5. Detections=[];
6. PlagPassages=[];
7. for all Pi ∈ InitPlagPass
8.    Oi=subset(Pi,M2);
9.    OrigPassages=findPassages(Oi,θg);
10.   if size(OrigPassages)>1
11.      for all Oj ∈ OrigPassages
12.         Pj=subset(Oj,M1)
13.         PlagPassages=PlagPassages ∪ Pj;
14.      endfor
15.   else PlagPassages=PlagPassages ∪ Pi;
16.   endif
17. Detections=Detections ∪
                [PlagPassages, OrigPassages];
18. endfor
19. return Detections;
```

تصویر ۲: الگوریتم مربوط به تشخیص مرزهای متنی [۱۱]

فراپردازش [۱۲]

روشی که مطرح شد مبتنی بر نمایش SWNG است و همهٔ کلماتی که متعلق به مجموعه ۵۰ کلمهٔ کلیدی انتخابی نباشند را حذف می‌کند. تشخیص های انجام شده باید بررسی شوند تا مشخص شود که شباهت میان متن سرقت شده شناسایی شده و متن اصلی، در هنگامی که کل متن مدنظر قرار گیرد بالا است. به

[۱۲] Post-Processing

علاوه ما به مکانیسمی نیاز داریم که موارد سرقت شناسایی شده را سه درجه شباهت با متن اصلی نمره دهی کند.

۲-۱-۵- روش تجزیه مقدار منفرد[۱۳]

تجزیه مقدار منفرد (SVD) در حقیقت یکی از مهم‌ترین ابزارها در بازیابی اطلاعات به صورت «شاخص گذاری معنای نهانی» است. همچنین این روش یکی از مناسب‌ترین ابزارها از نظر به‌کارگیری ماتریس‌های پراکنده می‌باشد.

فرض: فرض کنیم که A یک ماتریس r با رتبه m×n باشد . همچنین فرض کنیم که $\sigma_1 \geq \cdots \geq \sigma_r$ مقادیر آیگن ماتریس $\sqrt{AA^T}$ باشند. سپس ماتریس های قطری $U = (u_1, \ldots, u_r)$ و $V = (v_1, \ldots, v_r)$ که بردارهای ستونی آن‌ها قطری است و ماتریس $\Sigma = diag(\sigma_1, \ldots, \sigma_r)$ وجود دارند. تجزیه $A = U \Sigma V^T$ را تجزیه مقدار منفرد ماتریس A می‌نامند و اعداد $\sigma_1, \ldots, \sigma_r$ مقادیر منفرد ماتریس A هستند. ستون های U (یا V)، بردارهای چپ (یا راست) منفرد ماتریس A نامیده می‌شوند.

در اینجا ما تجزیه ماتریس اصلی A را داریم. نیازی به گفتن نیست که بردارهای منفرد چپ و راست، پراکنده نیستند. ما حداکثر r منفرد غیر صفر را در اختیار داریم که رتبهٔ r بعد کوچکتر در میان دو بعد ماتریس است. با این حال ما فضای زیادی را در حافظه از طریق ذخیره‌سازی ماتریس عبارت-در-سند بدین شکل اشکال می‌کنیم. خوشبختانه از آنجا که مقادیر منفرد معمولاً به سرعت پایین می‌آیند، ما می‌توانیم تنها k مورد از بزرگ‌ترین مقادیر منفرد را به همراه مختصات برداری منفرد متناظر آن‌ها انتخاب کنیم و یک تجزیه منفرد کاهش یافته به k از A را بوجود آوریم.

فرض کنیم که k ، $0 < k < r$ و تجزیه مقدار منفرد A را داریم

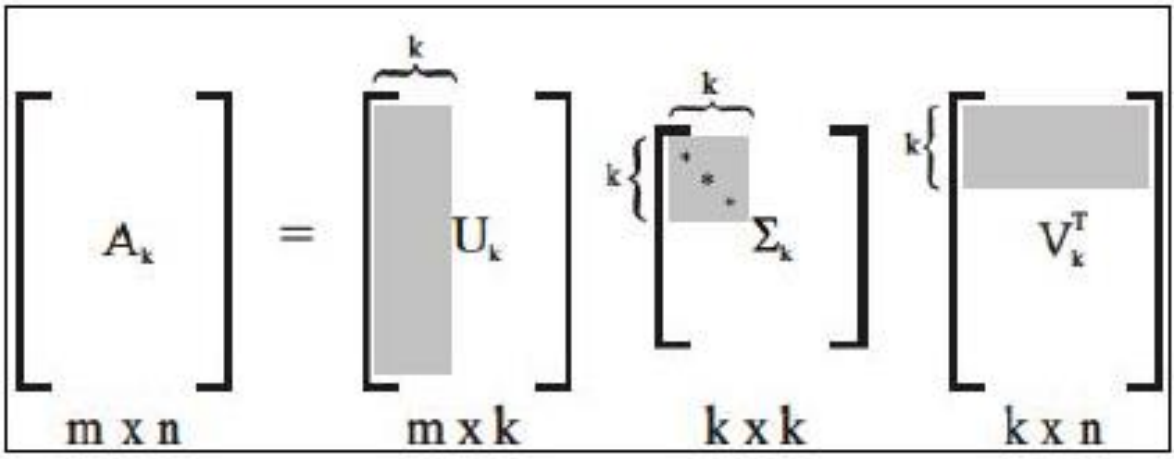
---

[۱۳] Singular Value Decomposition

$$A = U\sum V^T = (U_k U_0)\begin{pmatrix}\sum_k & 0\\ 0 & \sum_0\end{pmatrix}\begin{pmatrix}V_k^T\\ V_0^T\end{pmatrix}$$

ما $A_k = U_k \Sigma_k V_k^T$ را تجزیه مقدار منفرد کاهش یافته به k (*rank-k* SVD) می‌نامیم.

در بازیابی اطلاعات اگر هر سند مرتبط با یک موضوع باشد که ما می‌توانیم معانی نهان (latent semantics) را به دست آوریم- کلمات و اسناد مرتبط از نظر معنایی دارای بردارهای یکسانی در فضای کاهش یافته خواهند بود. برای نمایش تجزیه *rank-k* SVD ، تصویر ۳ را مشاهده کنید. نواحی خاکستری تعیین‌کننده اولین k مورد از مختصات بردارهای منفرد هستند که مورد استفاده قرار می‌گیرند.



تصویر ۳: تجزیه مقدار منفرد کاهش یافته به k [۱۱]

۲-۱-۶- نقشه های خود-آرایی (کوهوننیس)

یک نقشهٔ Kohonen را همچنین نقشه خود-آرایی[۱۴] (SOM) نیز می‌نامند که در سال ۱۹۸۹ توسط کوهوننیس به‌وجودآمد. [۹] این یک ابزار شبکه عصبی مصنوعی رقابتی است. SOM یکی از مؤثرترین ابزارها برای مشاهده بصری داده‌های چند بعدی تلقی می‌شود و همچنین مکانیسم مؤثری در پردازش سیگنال و کارکردهای استخراج داده ای است. SOM گونه ویژه‌ای از شبکه‌های عصبی است که می‌تواند یک سیگنال ورودی چند بعدی پیچیده را به یک سیگنال کم-بعدتر نگاشت یا ساده‌سازی کند. از آن برای دسته‌بندی

[۱۴] Self Organization Map

مجموعه داده‌ای با توجه به شباهت‌های مجموعه استفاده می‌شود. نقشه خود-آرایی یک شبکه مصنوعی است که در دو لایه نورون سازماندهی می‌شود. اولین لایه نشاندهنده داده‌های ورودی و دومین لایه نشان‌دهنده یک گرید نورون هستند ، که اغلب دوبعدی بوده و کاملاً یکدیگر مرتبط می‌باشند. همه گره‌های ورودی به همه نورون های (گره‌های) خروجی متصل می‌باشند. نورون های خروجی معمولاً در دو یا سه گرید کم-بعد آرایش می‌یابند. یک بردار وزنی برای هر نورون وجود دارد که دارای بعدیتی همانند بردارهای ورودی است. تعداد بعد های گرید خروجی معمولاً پایین‌تر از ابعاد ورودی است. SOMها اساساً برای کاهش بعدیت به کار گرفته می‌شوند و نه برای انبساط.

۲-۱-۷- ماتریس فاصله یکپارچه

ماتریس فاصله یکپارچه[۱۵] یا ماتریس U که نمایشی است از یک نقشهٔ خود-آرایی که فواصل میان نورون های شبکه و یا ماتریس‌های واحد را نشان می‌دهد، ابزار مؤثری برای مشاهده دسته‌ها در داده‌های ورودی بدون وجود هرگونه اطلاعات ما قبل در مورد دسته‌ها است. نمایش چند بعدی داده‌های نقشه‌ای خود-آرایی می‌تواند با استفاده از ماتریس فاصله یکپارچه انجام شود. این مسئله با استفاده از روابط توپولوژیکی میان نورون ها پس از فرایند یادگیری، حاصل می‌آید. ماتریس U دارای فواصل از هر مرکز واحد به همه همسایه‌های خود است. نورون های شبکه SOM در اینجا از طریق سلول‌های شش ضلعی نشان داده می‌شوند. با استفاده از ماتریس U ما می‌توانیم روابط توپولوژیکی میان نورون ها را شناسایی کنیم و در مورد ساختار داده ورودی استنباط‌هایی را انجام دهیم. از الگوهای رنگی برای نمایش شباهت‌ها استفاده می‌شود. رنگ‌های تیره (مقادیر بالا در ماتریس U) نشان دهنده آن هستند که فاصله‌ای میان مقادیر در فضای ورودی وجود دارد و مشخص می‌کند که هیچ شباهتی وجود ندارد. همچنین رنگهای روشن (مقادیر پایین در ماتریس U) نشان میدهند که بردار های در فضای ورودی به هم نزدیک هستند و این بیان‌کننده

[۱۵] Unified Distance Matrix

شباهت‌های زیاد در میان ورودی‌ها است. از این طریقه نمایش می‌توان برای بررسی ساختار فضای ورودی و دریافت دیدگاهی از ساختار ورودی غیر مشهود در فضای داده چند بعدی استفاده نمود.

مزایا و معایب: مهمترین مزایایی که این روش بدنبال آنها است عبارتند کاهش زمان جستجو در تشخیص سرقت ادبی و ساده سازی نمایش بصری و آنالیز نتایج.

۲-۱-۸- برچسب زنی نقش معنایی

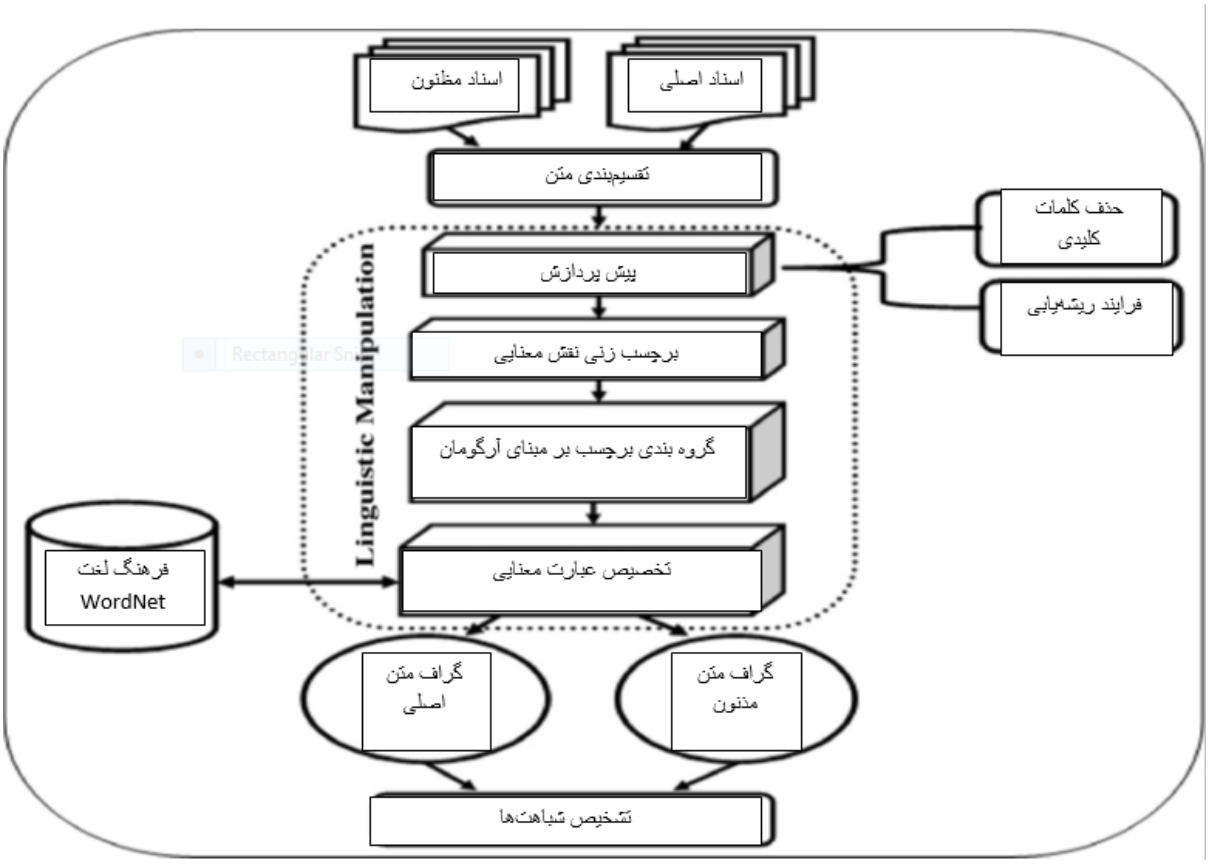
برچسب زنی نقش معنایی[۱۶] (SRL) [۷] یکی از روش های پردازش زبان طبیعی است که در حوزه های بسیاری مانند خلاصه سازی متن، دسته‌بندی متن و طبقه‌بندی متن کاربرد دارد. شیوه تشخیص سرقت ادبی بر مبنای SRL می‌تواند انواع سرقت ادبی شامل copy-paste، تغییر کلمات، جایگزینی مترادف ها ، تغییر ساختار کلمات در جمله ، ترکیب جملات از حالت غیر فعال به فعال و بالعکس را شناسایی کند. SRL برای ارزیابی معنای جملات به کار گرفته می‌شود و واژه نامه WordNet نیز برای استخراج مفاهیم یا مترادف ها برای هر کلمه در داخل جملات مورد استفاده قرار می‌گیرد. نمره وزنی برای هر آرگومان در جهت بررسی رفتار و تأثیرات آن‌ها در تشخیص سرقت ادبی محاسبه میشود.

در این روش ابتدا اسناد مظنون و اسناد اصلی با استفاده از تفکیک متن، خذف کلمات کلیدی، و ریشه‌یابی پیش پردازش می‌شوند. سپس از SRL برای تبدیل جملات به آرگومان ها بر مبنای موقعیت هر عبارت در جمله استفاده می‌شود. فعل‌های جملات نقش مهمی را در فرایند ایفا می‌کنند و آنالیز جملات مبتنی بر فعل‌های آن‌ها است. همه آرگومان های استخراج شده از متن، براساس نوع آرگومان در گره‌هایی گروه بندی می‌شوند. هر گره دارای آرگومان های استخراج شده مشابهی است و به واسطه نام آرگومان ، نام‌گذاری می‌شود، مانند Arg۰، Arg۱، V، Time ، Location... . این مرحله را «گروه بندی برچسب آرگومان» (ALG) می‌نامند. سپس ما همه مفاهیم هر عبارت را در گروه‌های آرگومان با استفاده از فرهنگ لغت WordNet استخراج می‌کنیم. این مرحله را تخصیص عبارت معنایی (STA) می‌نامیم. همه مفاهیمی که توسط فرهنگ

---

[۱۶] Semantic Role Labeling

لغت WordNet استخراج شدند، در یک گره با نام «گره موضوعی» گردآوری می‌شوند. مزیت یک گره موضوعی آن است که به سرعت می‌توانند بخش‌های مظنون را در اسناد پیدا کند. تصویر ۴ نشان دهنده ساختار عمومی برای این روش است.



تصویر ۴: ساختار روش SRL [۷]

۲-۱-۹- درهم سازی جملات

مزیت اصلی الگوریتم درهم سازی جملات[۱۷]،[۱۲] استفاده از روش درهم سازی برای جملات سند پردازش شونده است.

ابتدا هر سند به صورت تعدادی از جملات مجزا شده توسط یک نقطه مشاهده می‌شود. در حین پردازش تن، مکانیسم‌هایی وجود دارد که برای شناسایی مترادف و یا نقاط اعشار در اعداد به‌کارگرفته می‌شوند. این

[۱۷] Sentence Hashing

مسئله از آنجا اهمیت دارد که بسیاری از روش‌ها با نرمال سازی متن آغاز می‌شوند. عملکرد این الگوریتم به‌طورکلی تا حدی می‌تواند تحت تأثیر عدم نرمال سازی داده‌های ورودی قرار گیرد، و ذخیره سازی ساختار جمله برای موفقیت آن ضروری است. مسلماً متن می‌تواند دربرگیرنده تعدادی از بخش‌های ساختار بندی نشده مانند فهرست‌ها و لیست ها باشد. بدین ترتیب ایجاد فریم می‌تواند شرایطی که در آن نقطه ایدر پایان جملات وجود ندارد را مدیریت کند.

سپس، طولی از فریم درهم برای کل سیستم انتخاب می‌شود. فریم درهم از تعدادی از عبارات پی‌درپی حاصل از سند پردازش شده پدید می‌آید. طول این فریم بسیار اهمیت دارد زیرا در ارتباط مستقیم با عملکرد الگوریتم است. این فریم هدایت کننده ایجاد درهم های (hash) متعاقب می‌باشد. این فریم نباید بیش از حد کوتاه باشد زیرا هزینه‌های دفتری از جنبه‌های مثبت درهم سازی در مراحل بعدی الگوریتم، فراتر میروند ؛ و همچنین نباید بیش از حد بلند باشد زیرا بازدهی جستجو برای بلندترین چشمگیری کاهش پیدا می‌کند.

روش درهم سازی به واسطه دو پارامتر هدایت می‌شود: α و β. در هنگام پردازش متن، دنباله ای از عبارات به صورت یک فریم متنی منفرد در نظر گرفته می‌شود که باید در هنگام که طول کلی آن کمتر از مقدار پارامتر α باشد، در هم سازی شود، در غیر این صورت دنباله به تعدادی از فریم های متنی تقسیم میشود که طول آنها β است. خواه نتایج تخصیص دارای باقی مانده باشد یا خیر (که طول آن کمتر از مقدار α پس) آن را بر مبنای بخش اول قاعده پردازش می‌کنیم.

هر عبارت در فریم در یک عبارت منحصربه‌فرد نگاشت می‌شود، سپس مقدار مربوط به یک فریم معین به صورت مجموعه اعداد نشان دهنده آن فریم معین، محاسبه می‌شود. بدین ترتیب ، این مسئله باعث می‌شود که روش مورد نظر در مقابل ترتیب کلمات در عبارات مورد بررسی انعطاف‌پذیری داشته باشد، که یکی از ویژگی‌های بسیار مهم در کارکردهای تشخیص سرقت ادبی است. کل متن پردازش می‌شود تا جمله‌ای با اعداد نمایانگر یک سند یافت شود.

مثال زیر را با پارامترهای α برابر با ۹ و β برابر با ۱,۵ در نظر بگیرید. این مقادیر باید به گونه‌ای تفسیر شوند که جمله‌ای تا ۱۳ کلمه به صورت یک فریم متنی منفرد در نظر گرفته شود $(\alpha * \beta)$ و جملات بزرگ‌تر از آن به فریم هایی ۹-عبارتی افراز گردند.

On Tuesday Huntsman Corporation informed that it had revoked a bid worth $460 million for Rexene Corp. due to double rejection of its bids by the chemical company from Dallas.
On → 742 Tuesday → 110226 Huntsman → 289972 Corporation → 88818 informed → 57582 that → 66940 it → 956 had → 7532 revoked → 81758 a → 97
Frame hash: 704623
bid → 7212 worth → 161187 $460 → 22240 million → 33767 for → 7530 Rexene → 38381 Corp → 42432 due → 7437 to → 1039 double → 276024
Frame hash: 597249
rejection → 169746 of → 990 its → 7763 bids → 57811 by → 905 the → 8357 chemical → 337263 company → 354349 from → 60517 Dallas → 340253
Frame hash: 1337954

جدول ۱: مثال [۱۲]

مزایا و معایب: در روش درهم سازی، طول جملات یک مانع محسوب می شود. مهمترین کاری که می توان در این مورد انجام داد عبارت است از متراکم سازی معنایی برای کاهش بیشتر بردار مطرح شده.

۲-۱-۱۰- مدل و نمونه گیری

نکته کلیدی در این روش [۱۳] چگونگی اندازه‌گیری شباهت‌هاست. الگوریتم های قدیمی تر سرقت ادبی را از طریق انطباق رشته به واسطه مقایسه متن دو سند مشخص می‌کنند. برخی از ابزارها مبتنی بر این‌گونه الگوریتم ها هستند. محدودیت آن‌ها آنست که تنها می‌توانند متونی را بیابند که دقیقاً یکسان هستند و هیچ گونه تغییراتی در متن را نمی‌توانند کنترل کنند. یک روش بهینه عبارت است از محاسبه تناوب کلمات در سند و نمایش سند به صورت یک بردار مشخصه (ویژگی)، و سپس محاسبه شباهت از طریق معیار شباهت کسینوسی. این روش می‌تواند تشخیص سرقت ادبی در هنگام جابجا شدن لغات یا جملات و یا حذف یا افزوده شدن برخی کلمات انجام دهد. این گونه روش‌ها می‌توانند سرقت ادبی را با دقت بالاتری پیدا کنند. برخی از ابزارهای عملی با این روش‌ها ساخته می‌شوند.

در مقاله‌ای توسط ژانگ و فانگ، معیار شباهت کسینوسی[۱۸] از دو جنبه بهبود پیدا کرده است. جنبه اول پردازش سند است. پیش پردازش از آن جا لازم است که در یک سند ممکن است است کلمات جابه‌جا شده باشند و یا کلماتی با مفاهیم مشابه و یا کلمات کمکی استفاده شده باشند. [۱۴] پس از پیش پردازش، کلماتی با مفاهیم مشابه گروه بندی می‌شوند و کلمات کمکی از بردار مشخصه حذف می‌شوند. بهسازی دوم عبارت است از تنظیم وزن کلمات. اگر کلمه A در نوعی از اسناد مشابه کتابخانه به کار گرفته شده باشد، این به معنای سرقت ادبی نیست حتی اگر کلمه در هر دو سند چندین بار استفاده شده باشد. درغیراین‌صورت، اگر کلمه B با تناوب پایین در کتابخانه وجود داشته باشد، و در هر دو سند به طور متناوب استفاده شده باشد، سپس احتمال سرقت ادبی بسیار بالاتر است. بدین ترتیب، در بردار مشخصه، وزن کلمهٔ A باید کاهش پیدا کند و وزن کلمه B باید افزایش یابد.

مزایا و معایب: عدم پیش-پردازش صحیح سند باعث ایجاد خطا در نتایج می شود و وزن کلمات سند نیز باید مبتنی بر تناوب تکرار آنها در کتابخانه دیجیتال باشد زیرا در غیر این صورت شاخص شباهت کسینوس نمی تواند دقت کافی داشته باشد.

۲-۱-۱۱- متن تکراری

جملات دارای کلمات مشابه بین سندهای مظنون و منبع می‌توانند اولین نشانه سرقت ادبی به شمار روند. [۱۵] با این حال وجود آن‌ها به عنوان یک نشانه منحصربه‌فرد از سرقت ادبی نمی‌تواند چندان مورد اطمینان و اعتماد باشد، زیرا کاربری موضوعی کلمات نیز می‌تواند منجر به ایجاد جملاتی با کلمات مشابه (مثبت های کاذب) شود. به علاوه حتی تغییرات کوچکی برای مخدوش ساختن سرقت ادبی مانع از شناسایی جملات متناظر شده و منفی های کاذب پدید می‌آورد.

به منظور کنترل مشکلات بالا، سانچز وگا و همکاران یک استراتژی جدید را برای شناسایی یک متن سرقت شده با نام روش «شاخص بازنویسی» ارائه کرده است. [۱۳] این روش می‌تواند بخش‌های متن تکراری را

[۱۸] Cosine Similarity Measure

حتی در صورت ایجاد تغییرات در آن تشخیص دهد. به علاوه، هدف آنها تسهیل تشخیص سرقت ادبی به واسطه بررسی ویژگی‌های متنوع بخش‌های مربوط به متن تکراری درطی مرحله دسته‌بندی می‌باشد. روش شاخص بازنویسی می‌تواند یک وزن را به هر کلمه مربوط به سند مظنون اختصاص دهد تا درجهٔ عضویت آن‌ها در بخش متن سرقت شده نشان دهد. بدین ترتیب امکان کشف متنی که تغییراتی در آن اعمال شده است (مانند حذف لغات، جایگذاری لغات، و تغییر لغات) وجود دارد و این مسئله امکان یک انطباق جزیی را میان اسناد فراهم می‌آورد.

یک موضوع مهم دیگر که مورد توجه این محققین قرار گرفت، نمایش جامعتر بخش‌های متن تکراری[۱۹] است. این گونه نمایش به الگوریتم های دسته‌بندی امکان می‌دهد که تمایز بیشتری را میان اسناد سرقت شده و سرقت نشده از طریق ارائه ویژگی‌هایی که توصیف کننده تعداد بخش‌های متن تکراری و نیز اهمیت و دسته‌بندی آن‌ها می‌باشد، قائل شوند.

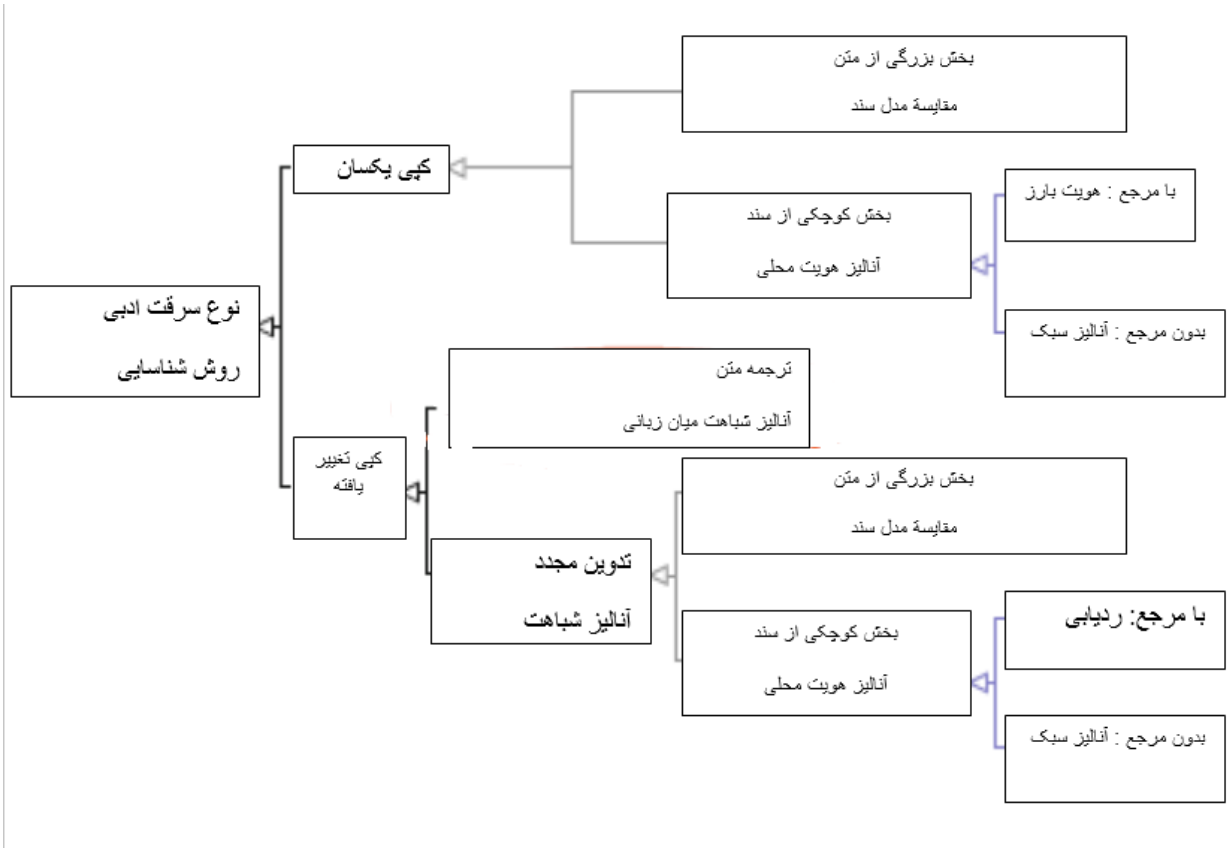
مزایا و معایب: این روش می تواند بخشهای متن تکراری را حتی در صورت قرار گرفتن تحت تغییرات مختلف شناسایی کند. همچنین در این روش تشخیص سرقت ادبی از طریق مد نظر قرار دادن ویژگی های مختلف بخشهای تکراری متن در طی مرحله دسته بندی ، تسهیل می شود .

[۱۹] Reused Text

فصل دوم : سرقت ادبی در زبان ها

۲-۲- سرقت ادبی برون-زبانی

سرقت ادبی میان زبانی در حقیقت مرتبط است با شناسایی اتوماتیک و استخراج سرقت ادبی در محیط‌های چند-زبانی. یک طبقه‌بندی از انواع سرقت ادبی و نیز روش‌های شناسایی آن‌ها در تصویر ۵ ارائه شده است.



تصویر ۵: طبقه بندی انواع سرقت ادبی متنی، به همراه روش‌های شناسایی آن‌ها[۱۶]

انواع روش‌های شناسایی سرقت ادبی برون زبانی

۲-۲-۱- سیستم‌های مبتنی بر لغت

این سیستم‌ها مبتنی هستند بر شباهت‌های لغات[۲۰] میان زبان‌ها (مثلاً انگلیسی-فرانسه) و تأثیرات زبانی (مثلاً computer در انگلیسی -> computadora در اسپانیایی) میان زبان‌ها. شباهت‌های لغات در زمان‌های مختلف می‌تواند در شکل‌گیری عبارات کوتاه منعکس شود؛ مثلاً پیشوند ها و یا nگرم های حرفی (کاراکتر). شاید دو مورد از اولین مدل‌های شباهت از این دست را بتوان «هم ریشه» بودن (بر اساس پیشوندها و سایر نشانه ها) و نمودار نقطه ای (بر مبنای ۴-گرم های کاراکتر) دانست. درحالی‌که این مدل‌ها اساساً برای هم‌تراز ساختن متون دوگانه پیشنهاد می‌شوند، اما آن‌ها برای تشخیص کاربری موجود تکراری در میان زبان‌ها (با برخی محدودیت‌ها) نیز کاربرد دارند.

۲-۲-۲- سیستم‌های مبتنی بر فرهنگ لغت

این سیستم‌ها لغات یا مفاهیمی مانند نام ماهیت هارا در یک فضای نمایش مشترک از طریق یک فرهنگ لغت[۲۱] چند زبانی نگاشت می‌کنند. با این حال فرهنگ‌های لغت چند زبانی همیشه در دسترس نیستند و همچنین Ceska دریافت که ناقص بودن فرهنگ لغت می‌تواند قابلیت‌های تشخیص را محدود کند.

۲-۲-۳- سیستم‌های مبتنی بر مجموعه‌های داده قابل مقایسه

این سیستم‌ها با استفاده از مجموعه‌های داده قابل مقایسه[۲۲] آموزش می‌بینند. یک نمونه از این امر را می‌توان آنالیز معنایی صریح میان زبانی دانست (CL-ESA). d_q و d' از طریق بردار شباهت‌ها با اسناد یک مجموعه شاخص CL، c_l، نمایش داده می‌شوند، یعنی $\overrightarrow{d_q} = \{sim(d_q, c_1), \ldots sim(d_q, c_l)\}, \overrightarrow{d'} = \{sim(d', c'_1), \ldots sim(d', c'_l)\} (c_i \in L, c'_i \in L')$ که

---

[۲۰] Lexicon

[۲۱] Thesaurus

[۲۲] Comparable Corpus

در آن sim یک مدل شباهت درون زبانی است، مانند معیار کسینوسی، و $\overrightarrow{d_q}$ و $\overrightarrow{d'}$ برای محاسبه $sim(d_q, d')$ مورد مقایسه قرار می‌گیرند.

۲-۲-۴- سیستم‌های مبتنی بر مجموعه‌های داده موازی

این سیستم‌ها با استفاده از مجموعه‌های داده ای موازی[۲۳]، برای یافتن تشابه های میان زبانی و یا یافتن مدل های ترجمه آموزش داده می‌شوند. اصول و منابع ماشین ترجمه (MT) در آن‌ها به کار گرفته می‌شود اما هیچ ترجمه حقیقی انجام نمی‌شود.

۲-۲-۵- سیستم‌های مبتنی بر ترجمه ماشینی

این مدل‌ها در CLPD رایج هستند و وظیفه مورد نظر را با تبدیل آن به یک مسئله تک-زبانی آسان‌تر می‌کنند. مسئله نمونه به صورت زیر است: ۱. یک شناساگر زبانی برای تعیین محتمل ترین زبان d_q به کار گرفته می‌شود؛ ۲. d_q در صورتی که به زبان مقایسه نوشته نشده باشد، ترجمه می‌شود؛ ۳. یک مقایسه درون زبانی بین d'_q و d' انجام می‌شود. [۸]

۲-۳- توصیف برخی از روش‌های مورد استفاده برای تشخیص سرقت ادبی برون زبانی

۲-۳-۱- ابراز فرت

Ferret ابزاری است برای شناسایی موجود مشابه متنی در مجموعه‌های بزرگی از اسناد. از آن با موفقیت در متون انگلیسی برای سال‌ها استفاده شده است. این یک ابزار رایگان و مستقل است که برای کاربران مبتدی و برای اجرا روی کامپیوترهای معمولی طراحی شده است و نتایج متوسطی را به دست می‌دهد. این ابزار دربرگیرنده تعداد زیادی از اسناد است ، مانند مقالاتی که توسط تعداد بسیاری از دانشجویان ارائه شده‌اند. همچنین از آن می‌توان برای شناسایی سرقت ادبی در کدهای کامپیوتری نیز استفاده کرد. بائو و همکارانش یک نسخه اصلاح شده از این ابزار را به کار گرفتند و مشاهده کردند که عملکرد آن در مورد

---

[۲۳] Parallel Corpus

سرقت ادبی در متون زبان چینی جالب توجه است. آثار دانشجویی از دو دانشگاه چین گردآور شدند و از Ferret برای تشخیص سرقت زدگی استفاده شد. نتایج به‌دست‌آمده توسط این محقق مشخص می‌کنند که Ferret می‌تواند سرقت ادبی مصنوعی و حقیقی (که پیش‌تر کشف نشده است) را شناسایی کند.[۱۷]

یک سیستم جالب توجه دیگر برای تشخیص کپی Turnitin است که از پایگاه داده بسیار بزرگی از مطالب روی وب و آثار پیشین دانشجویان استفاده می‌کند و آن‌ها را با آثار کنونی مقایسه می‌کند. با این حال اسناد باید برای پردازش در اختیار Turnitin قرار گیرند ، و یک هزینه تجاری برای آن وجود دارد. مقایسه‌ای از Ferret، Turnitin و سایر سیستم‌ها توسط Lyonارائه شده است. روش‌های دیگری وجود دارند که به شباهت‌های معنایی میان زوج‌های اسناد توجه می‌کنند . شخیص کپی برداری در کدهای کامپیوتری توسط این ابزار نیز موضوع بررسی‌های مالفولبود. [۱۷]

۲-۳-۲- طرح سیستم فرت

شناساگر کپی Ferret مجموعه‌ای از فایل‌ها را در اختیار می‌گیرد و شاخصی از شباهت را برای هر زوج از آن‌ها محاسبه می‌کند. مرحله اول، تبدیل یک سند به مجموعه‌ای از تری گرم های متداخل است. بدین ترتیب جمله‌ای مانند:

A storm is forcast for the morning

به مجموعهٔ تری گرم های زیر تبدیل می‌شود:

A storm is storm is forcast is forcast for

forcast for the for the morning

سپس مجموعه تری گرم ها برای هر سند با همه موارد دیگر مقایسه می‌شود و شاخص شباهت هر زوج از اسناد محاسبه می‌گردد. معمولاً نتایج به شکل یک جدول رتبه‌بندی شده ارائه می‌شوند که در آن مشابه ترین زوج‌ها در بالا قرار دارند.

سه استراتژی زیر برای پردازش رشته‌های کاراکتر هایی که یک جمله چینی را تشکیل می‌دهند برای انطباق Ferret با زبان چینی مورد استفاده قرار گرفتند:

استراتژی ابتدایی: کاراکترهای چینی همانند کاراکترهای انگلیسی مورد نظر قرار گرفتند؛ دنباله ها با مدنظر قرار دادن هر عنصری که کاراکتر چینی نیست (مانند فاصلهٔ، علایم نگارشی، اعداد و غیره) به عنوان یک مرز نشانه‌ای از یکدیگر تفکیک شدند .

استراتژی کاراکتر منفرد: به جای یافتن کلمات، کاراکترها به طور منفرد پردازش می‌شدند. هر کاراکتر منفرد در فایل متنی عنوان یک نشانه به کار گرفته می‌شود.

استراتژی فرهنگ لغت: بر مبنای یک فرهنگ لغت چینی، یک جمله به دنباله کلمات شناسایی شده در فرهنگ لغت تفکیک می‌شود.

۲-۳-۴ N-گرم

مدل n-گرم [۱۸] ابتدا در دسته‌بندی متون برمبنای اطلاعات آماری گرفته‌شده از کاربری دنباله کاراکترها مورد استفاده قرار گرفت. n-گرم ها در حقیقت کاراکترهای متداخل تناوبی شکل گرفته از یک جریان ورودی می‌باشند. آن‌ها را می‌توان عنوان جایگزینی برای بازیابی کلمه-محور متن مورد استفاده قرار داد.

محمد آ. خان و همکارانش یک سیستم تشخیص سرقت ادبی را برای موجود زبان اردو بر مبنای مدل n-گرم پیشنهاد کردند. آن‌ها از تری گرم برای نمایش متن بهره گرفتند. تری گرم بدان معناست که نشانه سه کلمه برای استخراج کلمات از متن مورد استفاده قرار می‌گیرد و این تری گرم ها با هم منطبق می‌شوند. سپس معیارهای انطباقی برای دسته‌بندی متن محاسبه می‌گردند. معیار شباهت R به‌صورت زیر تعریف می‌شود: [۱۸]

$$R = \frac{|S(A) \cap S(B)|}{|S(A) \cup S(B)|}$$

در زیر دو متن ۱ و ۲ مشاهده می‌شوند. متن ۱ در حقیقت متن اصلی (گرفته شده از یک کتاب به زبان اردو) است و متن ۲ نسخه دیگری از متن ۱ با تغییر کلمات است. تری گرم های مربوط به هر دو متن محاسبه می‌شوند؛ جدول ۲ فهرستی از تری گرم های محاسبه شده برای متن ۱ و جدول ۳ فهرستی از تری گرم های محاسبه شده برای متن ۲ را ارائه می‌کنند.

متن ۱:

برسات کا موسم شروع ہو چکا تھا۔ آسمان پر ہر وقت کالی کالی بدلیاں اٹکھیلیاں کرتی نظر آتیں اور ہلکی ہلکی بوندا باندی موسم کی خوبصورتی میں اضافہ کر دیتی۔ مگر ایک مسئلہ تھا۔

متن ۲:

برسات کا موسم جب شروع ہوتا ہے تو آسمان پر ہر طرف کالے بادل نظر آتے ہیں۔ اور ہلکی ہلکی پوھار موسم کو خوشگوار بنا دیتی ہے۔ مگر ایک مسئلہ ہے۔

برسات کا موسم	کا موسم شروع	موسم شروع ہو	شروع ہو چک	ہو چکا تھ	چکا تھا آسمان
تھا آسمان پر	آسمان پر ہر	پر ہر وقت	ہر وقت کالی	وقت کالی کالی	کالی کالی بدلیاں
کالی بدلیاں اٹکھیلیاں	بدلیاں اٹکھیلیاں کرتی	اٹکھیلیاں کرتی نظر	کرتی نظر آتیں	نظر آتیں اور	آتیں اور ہلکی
اور ہلکی ہلکی	ہلکی ہلکی بوندا	ہلکی بوندا باندی	بوندا باندی موسم	باندی موسم کی	موسم کی خوبصورتی
کی خوبصورتی میں	خوبصورتی میں اضافہ	میں اضافہ کر	اضافہ کر دیتی	کر دیتی مگر	دیتی مگر ایک
مگر ایک مسئلہ	ایک مسئلہ تھا				

جدول ۲: فهرست تری-گرم ها برای متن ۱ [۱۸]

برسات کا موسم	کا موسم جب	موسم جب شروع	جب شروع ہوتا	شروع ہوتا ہے	ہوتا ہے تو
ہے تو آسمان	تو آسمان پر	آسمان پر ہر	پر ہر طرف	ہر طرف کالے	طرف کالے بادل
کالے بادل نظر	بادل نظر آتے	نظر آتے ہیں	آتے ہیں اور	ہیں اور ہلکی	اور ہلکی ہلکی
ہلکی ہلکی پوھار	ہلکی پوھار موسم	پوھار موسم کو	موسم کو خوشگوار	کو خوشگوار بنا	خوشگوار بنا دیتی
بنا دیتی ہے	دیتی ہے مگر	ہے مگر ایک	مگر ایک مسئلہ	ایک مسئلہ ہے	

جدول ۳: فهرست تری-گرم ها برای متن ۲ [۱۸]

Passage	N	M	R	% Matching	Comments
J1	32	4	0.07	7%	Different
J2	29				
B1	48	26	0.3	30%	No Same
B2	44				
C1	64	6	0.04	4%	No Different
C2	74				
D1	42	10	0.133	13.3%	No
D2	43				
E1	40	19	0.3	30%	No Same
E2	45				
F1	44	37	0.73	73%	Yes Copied
F2	41				
G1	41	10	0.147	14.7%	No
G2	37				
H1	47	7	0.08	8%	Different
H2	38				
I 1	48	2	0.02	2%	Different
I 2	30				
A1	90	90	1	100%	Same
A2	90				

جدول ۴: نتایج مفید توصیف شده. ستون **Passage** شامل متن ها است، **N** تعداد کلی تری-گرم ها در متن است، **M** تعداد تری-گرم های منطبق می‌باشد، و **R** ضریب تشابه است. [۱۸]

۲-۵- خلاصه فصل دوم

سرقت ادبی به نام زبانی در حقیقت چیزی است که آن را به صورت دریافت عبارات، مفاهیم و جملات از متن دیگر و انتقال آن به متن خود در حیطه یک زبان منفرد تعریف می‌کنند. روش‌های بسیاری برای تشخیص سرقت ادبی توسط محققین در سال‌های گذشته به وجود آمده‌اند در اینجا به برخی از آن‌ها اشاره می‌کنیم که مرتبط با تشخیص ادبی در حیطه زبان منفرد هستند.

روش‌های مبتنی بر گرامر : روش مبتنی بر گرامر برای تشخیص سرقت ادبی است که بر ساختار گرامری اسناد تمرکز می‌کند و از یک شیوه انطباق رشته‌ای برای شناسایی و اندازه‌گیری شباهت میان اسناد بهره می‌گیرد. روش‌های مبتنی بر گرامر برای تشخیص کپی‌های دقیق بدون هرگونه تغییرات ، مورد استفاده قرار می‌گیرند اما نمی‌توانند متون کپی شده و در عین حال اصلاح شده را تشخیص دهند.

روش معنایی :روش معنایی مبتنی بر شباهت‌های میان اسناد به واسطه مدل فضای برداری می‌باشد. این روش همچنین می‌تواند حشویات کلمات را در سند شمرده و محاسبه کند و سپس از اثر انگشتهای اسناد دیگر برای یافتن شباهت استفاده کند.

روش ترکیبی گرامر-معنایی : این روش، که برای متون کپی شده، تغییر یافته، و بازنویسی شده ای مناسب است که نمی‌توان آن‌ها را از طریق روش های گرامر-محور و نیز روش‌های معنایی بررسی کرد.

سرقت ادبی برون-زبانی : سرقت ادبی برون زبانی در حقیقت مرتبط است با شناسایی اتوماتیک و استخراج سرقت ادبی در محیط‌های چند-زبانی. روشهای شناسایی سرقت ادبی برون زبانی عبارتند از:

سیستم‌های مبتنی بر لغت : این سیستم‌ها مبتنی هستند بر شباهت‌های لغات میان زبان‌ها (مثلاً انگلیسی-فرانسه) و تأثیرات زبانی (مثلاً computer در انگلیسی -> computadora در اسپانیایی) میان زبان‌ها. شباهت‌های لغات در زمان‌های مختلف می‌تواند در شکل‌گیری عبارات کوتاه منعکس شود؛ مثلاً پیشوند ها و یا nگرم های حرفی (کاراکتر). شاید دو مورد از اولین مدل‌های شباهت از این دست را بتوان «هم ریشه» بودن (بر اساس پیشوندها و سایر نشانه ها) و نمودار نقطه ای (بر مبنای ۴-گرم های کاراکتر) دانست. درحالی‌که این مدل‌ها اساساً برای هم‌تراز ساختن متون دوگانه پیشنهاد می‌شوند، اما آن‌ها برای تشخیص کاربری موجود تکراری در میان زبان‌ها (با برخی محدودیت‌ها) نیز کاربرد دارند.

سیستم‌های مبتنی بر فرهنگ لغت : این سیستم‌ها لغات یا مفاهیمی مانند نام ماهیت هارا در یک فضای نمایش مشترک از طریق یک فرهنگ لغت چند زبانی نگاشت می‌کنند. با این حال فرهنگ‌های لغت چند زبانی همیشه در دسترس نیستند و ناقص بودن فرهنگ لغت می‌تواند قابلیت‌های تشخیص را محدود کند.

سیستم‌های مبتنی بر مجموعه‌های داده قابل مقایسه : این سیستم‌ها با استفاده از مجموعه‌های داده قابل مقایسه آموزش می‌بینند. یک نمونه از این امر را می‌توان آنالیز معنایی صریح میان زبانی دانست (CL-ESA).

سیستم‌های مبتنی بر مجموعه‌های داده موازی : این سیستم‌ها با استفاده از مجموعه‌های داده ای موازی برای یافتن تشابه های میان زبانی و یا یافتن مدل های ترجمه آموزش داده می‌شوند. اصول و منابع ماشین ترجمه (MT) در آن‌ها به کار گرفته می‌شود اما هیچ ترجمه حقیقی انجام نمی‌شود.

سیستم‌های مبتنی بر ترجمه ماشینی : این مدل‌ها در CLPD رایج هستند و وظیفه مورد نظر را با تبدیل آن به یک مسئله تک-زبانی آسان‌تر می‌کنند.

روش	معایب	مزایا
روش گرامر-محور	نمی توانند متونی که شامل بازنویسی یا جابجایی کلمات هستند (اما معنای مشابهی دارند) را شناسایی نمایند.	روشهای گرامر-محور برای تشخیص کپی های دقیقی مناسب هستند که در آنها هیچ تغییری در متن مشاهده نمی شود؛
روش معنایی	چنانچه سند به صورت جزیی مورد سرقت قرار گرفته باشد، نمی‌تواند نتایج مناسبی را به دست دهد و این یکی از محدودیت‌های روش مذکور است.	روش معنایی مبتنی است بر سرقت ادبی غیر جزئی و از کل سند و نیز از فضای برداری برای انطباق میان اسناد استفاده می‌کند ؛
روش ترکیبی گرامر-معنایی	در این روش لازم است کلیه درستورات گرامی را وسیله ی ساختمان داده ای پیچیده ای معرفی گردد	برای تشخیص متون تغییر یافته از طریق بازنویسی یا جابجایی مناسب است که توسط روشهای گرامری صرف قابل تشخیص نیستند. همچنین این شیوه قادر است محدودیت های روش معنایی را نیز برطرف کند و موقعیت بخشهای سرقت شده را در سند تشخیص دهد.
Ferret	ارائه نتایج متوسط	این روش یک ابزار رایگان و مستقل و البته بسیار سریع است که کاربران مبتدی نیز می توانند از آن استفاده کنند.
N-گرم ها	پیچیدگی های استخراج و مقایسه آنها نیز ماکزیمم است.	تشخیص کپی توسط ۲-گرم ها ماکزیمم است.

جدول ۵ : مزایا ومعایب روش های تشخیص سرقت ادبی درون زبانی [۱۰]

فصل سوم : سرقت ادبی در کدهای کامپیوتری

۳-۱- سرقت ادبی در علوم کامپیوتر

سرقت ادبی در فایل‌های کد منبع زمانی اتفاق می‌افتد که کد منبع بدون اشاره به مالک و نویسندهٔ اصلی آن کپی برداری و ویرایش شود. براساس گفته جوی و لاک ، تغییرات لغوی در حقیقت تغییراتی هستند که می‌توانند در کد منبع بدون اثرگذاری بر تجزیه (parsing) برنامه انجام شوند و بر تجزیه کد اثر می‌گذارند و شامل باگ زدایی از برنامه میباشد. نمونه های تغییرات لغوی شامل تغییر نام شناساگرها و کامنت ها هستند، و از نمونه‌های تغییرات ساختاری نیز می‌توان به بازآرایی و جایگزینی جمله‌ها اشاره کرد. [۲]

سرقت ادبی کد منبع را میتوان بصورت انتقال بخشی از کد منبع نوشته شده توسط فردی دیگر به کد شخصی (بدون اشاره به اینکه کدام بخش‌ها از محقق دیگری کپی شده اند) تعریف کرد.

سرقت ادبی معمولاً در محیط‌های آکادمیک اتفاق می‌افتد. دانشجویان به صورت عمدی یا غیرعمدی بخشی از منبع را بدون ذکر نام، در اثر خود می‌گنجانند. تشخیص سرقت ادبی به صورت دستی در مجموعه‌ای از صدها اثر غیرممکن و غیراقتصادی است. بدین ترتیب، ابزارهای نرم‌افزاری برای کمک به مدرسین در تشخیص سرقت ادبی به‌وجود آمده‌اند.

باید دانست که هیچ یک از این ابزارها نمی‌توانند حقیقتاً وجود یا عدم وجود سرقت ادبی را نشان دهند. این ابزارها صرفاً یک شاخص مشابهت را برای هر زوج از برنامه‌ها به دست می‌دهند و سپس بررسی‌های انسانی برای تعیین این مسئله لازم است که این شباهت‌ها در حقیقت سرقت ادبی هستند یا آنکه صرفاً به واسطه انجام یک تکلیف به صورت استاندارد و عرف مشاهده شده‌اند.

۳-۲- مطالب سرقت شده در کدها

سرقت ادبی در تکالیف برنامه‌نویسی می‌تواند فراتر از کپی کردن کد منبع[۲۴] باشد؛ این کار ممکن است دربرگیرنده کامنت ها، داده‌های ورودی برنامه، و طرح اینترفیس باشد. [۲]

فرآیند بررسی اینکه آیا سرقت ادبی در کد منبع انجام شده است یا خیر، می‌تواند شامل بررسی بخش‌های دیگری از یک تکلیف برنامه‌نویسی باشد زیرا در برخی از موارد، کد منبع به‌تنهایی برای شناسایی و اثبات سرقت ادبی کافی نیست. یک تکلیف برنامه‌نویسی می‌تواند شامل نمودارهای طراحی، کد منبع و سایر اسناد باشد. از افراد آکادمیک پرسیده شد که در هر یک از سناریو های زیر، یکی از پاسخ‌های «موافق»، «مخالف»، و «نه موافق نه مخالف»، را انتخاب کنند.

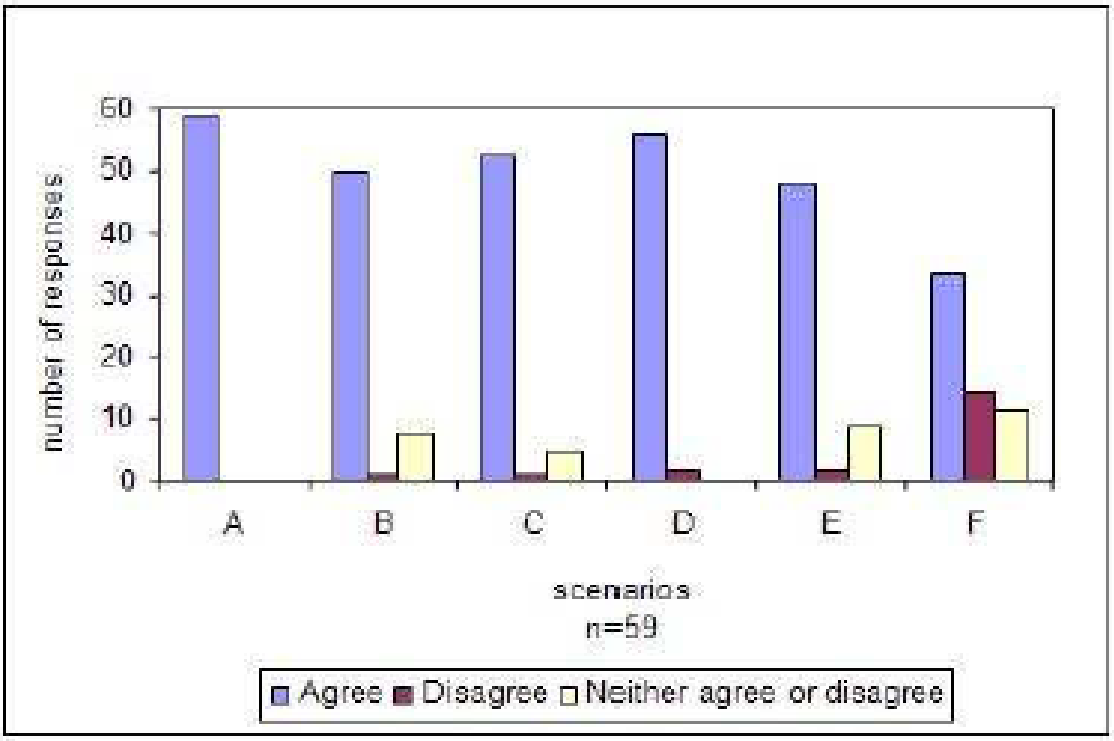
- سرقت ادبی در تکالیف برنامه‌نویسی می‌تواند شامل موارد زیر باشد:
  - سناریو A: کد منبع یک برنامهٔ کامپیوتری
  - سناریو B: کامنت های[۲۵] موجود در کد منبع
  - سناریو C: مطالب طراحی یک برنامهٔ کامپیوتری
  - سناریو D: سند یک برنامهٔ کامپیوتری
  - سناریو E: اینترفیس کاربری یک برنامهٔ کامپیوتری
  - سناریو F: داده‌های ورودی برنامه، برای آزمایش برنامه

پاسخ‌های ارائه شده در تصویر ۶ آمده‌اند.

---

[۲۴] Source Code

[۲۵] Comment



تصویر ۶: سناریو ها و پاسخ‌ها [۲]

همه افراد موافق بودند که در یک تکلیف برنامه‌نویسی، کد منبع می‌تواند مورد سرقت ادبی قرار گیرد. کامنت های موجود در کد منبع نیز می‌تواند سرقت شود و البته می‌تواند به شناسایی موارد سرقت کد منبع کمک کند. همچنین اغلب افراد توافق داشتند که کامنت ها نیز قابل سرقت هستند و همچنین ممکن است بتوانند به شناسایی موارد سرقت کد منبع کمک کنند.

داده‌های ورودی برنامه و اینترفیس کاربری نیز در صورتی که بخشی از الزامات تکلیف مورد نظر باشند، ممکن است مورد سرقت قرار گیرند. اغلب پاسخ‌دهندگان موافق بودند که داده‌های ورودی برنامه می‌توانند سرقت شوند، اما این مسئله به تنهایی به شناسایی سرقت ادبی کمکی نمی‌کند. سه نفر از افراد آکادمیک اعتقاد داشتند که کپی کردن داده‌های ورودی در هنگامی مسئله ساز می‌شود که دانشجویان از نظر استراتژی‌های آزمایشی مورد سنجش قرار گیرند. هنگامی که این کار انجام می‌شود، ارزیابی سرقت ادبی با مشاهدهٔ استراتژی آزمایشی، شامل پایگاه‌های داده‌ای (مثلاً داده‌های ورودی) مورد استفاده برای آزمودن

برنامه ، و مطالب آزمایشی، شامل طرح آزمایش، سند طراحی سیستم، سند فنی و راهنماهای کاربری میسر خواهد بود.

طرح‌های مذنون مربوط به اینترفیس که توسط دانشجویان ارائه می‌گردند نیز باید از نظر سرقت مورد بررسی قرار گیرند. افراد عنوان کردند که این که یک اینترفیس و کاربری در معرض سرقت قرار گیرد یا خیر مبتنی بر این مسئله است که آیا در تکلیف مورد نظر ، طراحی اینترفیس از دانشجو خواسته شده است یا خیر.

۳-۳- انواع سیستم‌های شناسایی سرقت ادبی

سیستم‌های شناسایی سرقت ادبی را می‌توان به انواع هرمتیک و وب ، و از جنبه‌ای دیگر به انواع چندمنظوره، زبان طبیعی و کد منبع تقسیم‌بندی کرد. سیستم‌های شناسایی مربوط به وب تلاش میکنند که موارد انطباق مربوط به اسناد مذنون در منابع آن‌لاین را بیابند. سیستم‌های هرمتیک تنها در یک مجموعه محلی از اسناد، به دنبال نمونه‌های سرقت ادبی می‌گردند. این سیستم‌ها دارای یک پایگاه داده‌ای از اسناد هستند. برای مثال، پایگاه داده‌ای ممکن است دربرگیرنده آثار سایر محققین و نیز مطالب مورد استفاده در یک واحد درسی خاص باشد.

۳-۴- ابزارهای شناسایی سرقت ادبی کد منبع

ابزارهای مختلفی برای شناسایی سرقت ادبی وجود دارند که می‌توانند بر مبنای الگوریتم، تقسیم بندی شوند. در این بخش به توصیف ابزارهای شناسایی سرقت ادبی کد منبع و با استفاده از دسته‌بندی‌های معرفی شده توسط موزگووی می‌پردازیم. آن‌ها شامل سیستم‌های مبتنی بر فینگرپرینت ، و روش‌های مقایسه محتویات هستند. دسته‌بندی‌های مختلف دیگری نیز در مقالات وجود دارند.

۳-۴-۱- سیستم‌های مبتنی بر اثر انگشت (فینگرپرینت)

در سال‌های ابتدایی، فینگرپرینت ها در سیستم‌های «شمارش مشخصه» به‌کارگرفته می‌شدند. اولین سیستم شناخته شده تشخیص سرقت ادبی یک برنامه شمارش مشخصه بود که توسط اوتن اشتاین برای شناسایی آثار یکسان یا نزدیک به هم مربوط به دانشجویان پدید آمد. این برنامه از شاخص‌های نرم‌افزاری هالستد

برای شناسایی سرقت ادبی به واسطه شمارش عملگرها و عملوندها (operator and operand) برای مدول های ANTI-FORTRAN استفاده کرد. شاخص‌های به کار گرفته شده توسط هالستد عبارت بودند از: [۲]

- تعداد عملگرهای[۲۶] منحصربه‌فرد
- تعداد عملوندهای منحصربه‌فرد
- تعداد کلی وقوع عملگر ها
- تعداد کلی وقوع عملوندها[۲۷]

رابینسون و سوفا یک برنامه تشخیص سرقت ادبی را به کار گرفتند که ترکیبی بود از شاخص‌های جدید و شاخص‌های هالستد به منظور بهینه‌سازی شناسایی سرقت ادبی. سیستم آن‌ها با نام ITPAD متشکل از سه مرحله بود: آنالیز لغوی، آنالیز ساختار برنامه برای ویژگی‌ها، و آنالیز ویژگی‌ها. ITPAD (ابزار سازمانی برای اصلاح برنامه) هر برنامه را به چند بلوک تقسیم می‌کند و یک گراف را برای نمایش ساختار برنامه می‌سازد. سپس فهرستی از ویژگی‌ها را بر مبنای آن لیست‌های لغوی و ساختاری پدید می‌آورد و زوج‌های برنامه‌ها را با شمارش این ویژگی‌ها مقایسه می‌کند. [۲]

رامبالی و سیج یک سیستم شمارش مشخصه را پدید آوردند که برنامه‌های دانشجویان را می‌گیرد، آن‌ها را تفکیک (parse) می‌کند و سپس یک «سیستم اطلاعات» را می‌سازد که دربرگیرنده بردارهای دانش است و در آن هر بردار اطلاعاتی در مورد مشخص ها در برنامه یک دانشجو را در اختیار می‌گیرد. این محققین از ویژگی‌های مشابه شناسایی شده توسط سیستم‌های پیش‌گفته استفاده می‌کنند، با این حال از روش مجزایی بهره می‌برند که ویژگی‌های حلقه-ساز [۲۸] را در یک برنامه می‌شمارد. به جای شمارش گونه‌های جملات حلقه-ساز به صورت مجزا، آن‌ها در میان انواع حلقه‌ها تفکیک قائل نمی‌شوند، و شمارش همه این جملات را در شمارش یک مشخصه در نظر می‌گیرند. هنگامی که بنیان‌های دانش ایجاد می‌شوند، برنامه‌ها در یک

---

[۲۶] Operator
[۲۷] Operand
[۲۸] Loop Building

«درخت تصمیم» دسته بندی می‌گردند. بر مبنای درخت تصمیم، برنامه‌هایی که حاوی شباهت‌هایی هستند، شناسایی خواهند شد. [۲]

از آن زمان، ابزارهای پیشرفته‌ترین برای شناسایی سرقت ادبی به وجود آمدند. این ابزارها در مقالات عموماً با عنوان «سیستم‌های شاخص ساختاری» شناخته می‌شوند. سیستم‌های ساختاری مقایسه‌ای را در مورد ساختار برنامه‌ها از نظر شباهت انجام دهند. موزگووی آن‌ها را عنوان «روش‌های مقایسه محتویات» دسته بندی کرده است. [۲]

مقایسه سیستم‌های مبتنی بر انطباق رشته و شمارش مشخصه نشان داده است که روش‌های شمارش مشخصه به‌تنهایی برای شناسایی سرقت ادبی کافی نیستند. ابزارهای جدید در شناسایی سرقت ادبی مانند MOSS ، ترکیبی از روش‌های فینگرپرینت با روش شاخص ساختاری است.

MOSS (معیار شباهت نرم‌افزاری) در سال ۱۹۹۴ توسط آیکن و برکلی بعنوان سیستمی برای بررسی شباهت های کد منبع نوشته شده در C، C++ جاوا و پاسکال بوجود آمد مبتنی است و یک الگوریتم انطباق رشته که به واسطه تقسیم برنامه به k-گرم ها عمل می‌کند، که در آن k-گرم یک زیر-رشته پیوسته با طول k است. هر k-گرم درهم (hash) می‌شود و MOSS زیرمجموعه از این مقادیر درهم را به صورت فینگرپرینت برنامه انتخاب می‌کند. شباهت به واسطه تعداد فینگرپرینت های مشترک در برنامه‌ها تعیین می‌شود؛ یعنی هرچه فینگرپرینت های مشترک بیشتری وجود داشته باشد، آن‌ها مشابه تر هستند. برای هر زوج از قطعه های کد منبع شناسایی شده، خلاصه نتایج شامل تعداد نشانه‌های (token) منطبق، و در صد تداخل کد منبع بین زوج های فایل ها است. [۱۹].

روش‌های مقایسه محتویات معمولاً با عنوان سیستم‌های شاخص ساختاری در مقالات شناخته می‌شوند. این سیستم‌ها، برنامه‌ها را به نشانه‌ها تبدیل می‌کنند و سپس دنباله ای پیوسته و منطبق از زیر-رشته‌های موجود در برنامه‌ها را جستجو می‌کنند. شباهت میان برنامه‌ها مبتنی است بر در صد متن‌های منطبق.

موزگووی روش‌های مقایسه محتویات را به الگوریتم های مبتنی بر انطباق رشته، الگوریتم‌های انطباق پارامتری سازی شده، و الگوریتم های مقایسه درخت‌های تجزیه (parse trees) تقسیم می‌کند.

۳-۴-۲- الگوریتم های انطباق رشته[۲۹]

جدید ترین سیستم‌های تشخیص سرقت ادبی متکی هستند بر مقایسه ساختار برنامه‌ها. این سیستم‌ها از شاخص‌های شمارش مشخصه استفاده می‌کنند اما همچنین ساختار برنامه را نیز به منظور بهبود فرآیند تشخیص سرقت ادبی، مقایسه می‌نمایند. سیستم‌های مبتنی بر انطباق رشته از الگوریتم های مقایسه‌ای استفاده می‌کنند که پیچیده‌تر از الگوریتم های شمارش مشخصه هستند. اغلب الگوریتم های انطباق رشته واسطه تبدیل الگوریتم ها به نشانه‌ها عمل می‌کنند و سپس از یک الگوریتم پیچیده جستجو برای شناسایی زیر-رشته‌های مشترک متنی بین و برنامهٔ استفاده می‌نمایند.

این سیستم‌ها ابتدا توسط دونالدسون پیشنهاد شدند. این محقق روش‌های ساده‌ای را شناسایی کرد که دانشجویان تازه کار برنامه‌نویسی برای تشخیص سرقت ادبی از آن‌ها استفاده می‌کنند. [۱۹]. این روش‌ها عبارتند از:

- تغییر نام متغیرها
- بازآرایی جملاتی که بر نتیجه برنامه اثر گذار نیستند
- تغییر فرمت جملات
- شکستن جملات، مثلاً تعریف‌های چندگانه و جملات خروجی

روش‌های شناسایی شده توسط دونالدسون اساسی‌ترین اشکال حمله هستند. ویل و جوی و لاک نیز فهرست دقیقی از حملات را ارائه کرده‌اند. یک فهرست بسیار کامل‌تر توسط پرچلت مطرح شد. ما همچنین یک فهرست جامع را که دربرگیرنده حملات سرقت ادبی هستند ارائه می‌کنیم.[۲]

---

[۲۹] String Matching

برنامه ارائه شده توسط دونالدسون به اسکن فایل‌های کد منبع می‌پردازد و اطلاعات مربوط به انواع خاصی از جملات را ذخیره سازی می‌کند. سپس یک کاراکتر منفرد کد به انواع جملاتی که در توصیف ساختار اهمیت دارند، اختصاص می‌یابند. سپس هر تخصیص به صورت رشته‌ای از کاراکترها نمایش داده می‌شود. اگر نمایش رشته‌ها یکسان یا مشابه باشد، زوج برنامه‌های مورد بررسی یکسان تلقی می‌شوند.

برخی از سیستم‌های جدید مبتنی بر انطباق رشته شامل Plague، Yap۳ (که خود نوع دیگری از Plague است)، JPlag، و Sherlock هستند [۲۰].

در اغلب سیستم‌های مبتنی بر انطباق رشته، جمله موارد پیش گفته، مرحله اول نشانه سازی (tokenization) نامیده می‌شود. در مرحله نشان سازی، هر فایل کد منبع از نشانه‌های از پیش تعیین‌شده و منسجم جایگزین می‌شود؛ مثلاً انواع مختلف حلقه‌ها در کد منبع می‌تواند با نام نشانه مشابهی از همان نوع حلقه جایگزین شود (مثلاً حلقه while، حلقه for). پس هر فایل کد منبع نشانه سازی شده با یک سری رشته‌های نشانه‌ای نمایش داده می‌شود. سپس برنامه‌ها از طریق جستجو برای دنباله های زیر- رشته ای انطباقی مربوط به نشانه‌ها مقایسه می‌شوند. برنامه‌هایی که در برگیرنده تعدادی نشانه‌های انطباقی در بالای یک آستانه معین باشند، یکسان تلقی می‌شوند. شباهت میان دو فایل اساساً به واسطهٔ پوشش نشانه‌های انطباقی میان فایل‌های شناسایی شده محاسبه می‌شود.

Plague ابتدا ترتیبی از نشانه‌ها را برای هر فایل ایجاد می‌کند و سپس نسخه‌های نشانه‌ای برنامه‌های انتخابی را با استفاده از یک روش انطباق رشته ، مقایسه می‌نماید. نتایج به صورت فهرستی از زوج‌های انطباقی نمایش داده می‌شوند که با ترتیب درجه مشابهت مربوط به طول بخش انطباقی از دنباله های نشانه میان دو فایل ارائه شده‌اند. با این حال مشکلاتی در زمینه Plague مطرح گشته‌اند. یکی از مشکلات Plague آن است که تبدیل آن به یک زبان برنامه‌نویسی دیگر وقت‌گیر است. بعلاوه، نتایج در دو فهرستی نشان داده می‌شوند که به واسطه شاخص‌هایی مرتب شده‌اند که تفسیر نتایج را مشهود نمی‌سازند. در آخر اینکه Plague چندان

بازدهی بالایی ندارد زیرا مبتنی بر تعدادی از ابزارهای الحاقی Unix که باعث ایجاد مشکلاتی در قابلیت انتقال می‌شود.

YAP۳ برنامه‌ها را به رشته‌ای از نشانه‌ها تبدیل می‌کند و آن‌ها را با استفاده از یک الگوریتم انطباق نشانه ، الگوریتم RKR-GST به منظور یافتن قطعات مشابه کد منبع ، مقایسه می‌نماید. YAP۳ فایل‌های کد منبع را پیش از تبدیل به نشانه‌ها، پیش پردازش می‌کند. این پیش پردازش شامل حذف کامنت ها، تبدیل حروف بزرگ به حروف کوچک، نگاشت مترادف ها به یک شکل مشترک (یعنی تابع به روش نگاشت می‌شود)، بازآرایی توابع در ترتیب فراخوانی آن‌ها، و حذف همه نشانه‌هایی که مربوط به لغات زبان هدف نیستند (یعنی حذف همه عباراتی که جز عبارات محفوظ برای زبان نیستند). YAP۳ اساساً برای تشخیص شکستن توابع کد به توابع متعدد ، و نیز برای شناسایی بازآرایی قطعه‌های مستقل کد منبع بوجود آمد. این الگوریتم به واسطه مقایسه دو رشته (الگو و متن) کار می‌کند که دربرگیرنده جستجوی متنی برای یافتن زیر-رشته‌های انطباقی الگو می‌باشد. بخش‌های انطباقی زیر-رشته‌ها، مجموعه معنی دار[۳۰] نامیده می‌شوند. هر مجموعه معنی دار در حقیقت یک انطباق است که دربرگیرندهٔ یک زیر-رشته از الگو و یک زیر-رشته از متن می‌باشد. هنگامی که انطباقی یافت شد، موقعیت نشانه‌های مجموعه معنی دار ها تنظیم می‌شود. از مجموعه معنی دار هایی که طول آن‌ها در زیر یک آستانه طول انطباق مینیمم است صرف‌نظر می‌شود. هدف از الگوریتم RKR-GST یافتن انطباقی های ماکسیمال دنباله های زیر-رشته پیوسته است که دربرگیرنده نشانه‌هایی است که توسط زیر-رشته‌های دیگر پوشش داده نشده است و در نتیجه تعداد نشانه‌های پوشش داده شده توسط مجموعه معنی دار ها به حداکثر می‌رسد.

JPlag نیز از الگوریتم مقایسه ای مشابه YAP۳ استفاده کند اما با بازدهی زمان اجرای بهینه. در JPlag، مشابهت به واسطه در صد رشته‌های نشانه ای پوشش داده شده محاسبه می‌شود. یکی از مشکلات JPlag آنست که فایل‌ها باید تجزیه شوند تا در مقایسه مربوط به سرقت ادبی گنجانده شوند، و این مسئله باعث از

---

[۳۰] tile

دست رفتن فایل‌های مشابه می‌شود. همچنین پارامتر طول انطباق مینیمم که در JPlag توسط کاربر تعریف می‌شود، در یک عدد پیش‌فرض قرار داده می‌شود. تغییر این عدد می‌تواند نتایج شناسایی را بهتر یا بدتر کند، و برای تغییر این عدد، نیاز به درکی از الگوریتم ماورای JPlag وجود دارد (یعنی RKR-GST). JPlag به صورت یک سرویس وب اجرا می‌شود و در برگیرنده یک اینترفیس کاربری ساده اما مؤثر است. اینترفیس کاربری نشاندهنده فهرستی از زوج‌های فایلی مشابه و درجه مشابهت آن‌ها است، و همچنین شامل نمایش مقایسه‌ای از فایل‌های مشابه شناسایی شده بواسطه برجسته سازی بلوک‌های انطباقی آن‌ها از قطعات کد منبع می‌باشد.

Sherlock نیز یک الگوریتم مشابه YAP۳ را اجرا می‌کند. Sherlock برنامه‌ها را به نشانه‌ها تبدیل می‌کند و دنبال های خطوطی را (که run) نامیده می‌شوند جستجو می‌کند که در دو فایل مشترک هستند. همانند الگوریتم YAP۳ ، Sherlock نیز به دنبال runهای با طول مشابه می‌گردد. اینترفیس کاربری Sherlock فهرستی از زوج‌های فایل‌های مشابه و درجه شباهت آن‌ها را نمایش می‌دهد و بلوک‌های انطباقی قطعه‌های کد منبع آن‌ها را که در زوج فایل‌های شناسایی شده یافت شده‌اند، مشخص می‌کند. به‌علاوه، Sherlock نمایش سریعی از نتایج را به شکل یک گراف ایجاد می‌کند که در آن هر رأس بیان‌کننده یک فایل منفرد کد منبع و هر گوشه نشان‌دهنده درجه شباهت میان دو فایل است. گراف تنها شباهت (گوشه‌های) میان فایل‌ها در بالای یک سطح تعریف شده توسط کاربر می‌باشد. یکی از مزایای Sherlock آن است که برخلاف JPlag، نیازی به تجزیه فایل‌ها به منظور گنجاندن آن‌ها در مقایسه وجود ندارد و هیچ پارامتر تعریف شده توسط کاربر ، که بر عملکرد سیستم اثرگذار باشد، موجود نیست. Sherlock یک ابزار منبع باز است و روش انطباقی نشانهٔ آن به‌راحتی قابل هماهنگ سازی با زبان‌های دیگری به جز Java می‌باشد. Sherlock ابزار مستقل است و یک سرویس مبتنی بر وب مانند JPlag و MOSS نیست. یک ابزار مستقل می‌تواند برای افراد آکادمیک از نظر بررسی فایل‌های دانشجویی به دنبال سرقت ادبی ، با مدنظر قرار دادن مسائل محرمانگی ، مطلوب‌تر باشد.

Plaggie ابزاری است همانند JPlag اما بدون الگوریتم‌های بهینه‌سازی سرعت. Plaggie فایل‌ها را به نشانه‌ها تبدیل می‌کند و از الگوریتم RKR-GST برای مقایسه فایل‌ها از نظر شباهت استفاده می‌کند. ایده ماورای

Plaggie ایجاد ابزار همانند JPlag است که بتواند مستقل باشد (یعنی بتواند روی یک دستگاه محلی نصب شود) و یک ابزار مبتنی بر وب نباشد ، و اینکه دارای کارکردهای بیشتری بوده و افراد آکادمیک را قادر سازد که کد را از طریق مقایسه استخراج کنند.

یک ابزار دیگر شناسایی سرقت ادبی که توسط موزگووی مطرح شد، سیستم تشخیص سریع سرقت ادبی (**FDPS**) بود که هدف آن بهسازی سرعت تشخیص سرقت ادبی با استفاده از یک ساختار داده شاخص گذاری شده برای ذخیره‌سازی فایل‌ها بود. در ابتدا فایل‌ها به نشانه‌ها تبدیل می‌شوند و فایل‌های نشانه گذاری شده در یک ساختار داده‌ای شاخص بندی شده قرار می‌گیرند. این مسئله امکان جستجوی سریع فایل‌ها را با استفاده از الگوریتمی مشابه الگوریتم مورد استفاده در YAP۳ فراهم می‌آورد. این کار شامل اتخاذ یک «فایل آزمایشی» و جستجو بدنبال زیر-رشته انطباقی در «فایل مجموعه» میباشد. موارد انطباقی در یک مخزن ذخیره سازی می‌شوند و سپس برای محاسبهٔ شباهت‌های میان فایل‌ها مورد استفاده قرار می‌گیرند. شباهت عبارتست از نسبت تعداد کلیه نشانه‌های منطبق در فایل مجموعه به تعداد کلیه نشانه‌ها در فایل آزمایشی. این نسبت می‌تواند پوشش (جامعیت) نشانه‌های منطبق با تعیین کند. زوج های فایل‌هایی دارای یک مقدار شباهت بالای یک حد آستانه معین بازیابی می‌شوند. یکی از مشکلات ابزار FDPS آن است که نتایج آن را نمی‌توان به واسطه نمایش قطعه های کدهای مشابه مشاهده کرد. بدین ترتیب، محققین FDPS را با ابزار Plaggie ترکیب کردند تا مقایسه فایل به فایل را انجام داده و نتایج را مشاهده کنند.[۲۰]

۳-۴-۳- الگوریتم های انطباقی پارامتری سازی شده

الگوریتم های انطباق پارامتری سازی[۳۱] شده همانند روش‌های متداول انطباق نشانه هستند، اما نشانه ساز پیشرفته‌تری دارند. اساساً این الگوریتم های انطباق پارامتری سازی شده با تبدیل فایل‌ها به نشانه‌ها کار می‌کنند. یک انطباق پارامتری (که آن را انطباق- p نیز می‌نامند)، قطعه‌های کد منبعی را که نام های متغیر آن‌ها به‌صورت سیستماتیک جایگزین شده است (تغییر نام) را منطبق می‌نمایند.

---

[۳۱] Parameterized Matching Algorithm

ابزار **Dup** مبتنی است بر یک الگوریتم انطباق-p و برای شناسایی کد کپی شده در نرم‌افزار ارائه شده است. این ابزار می‌تواند بخش‌های یکسان و پارامتری سازی شده کد منبع را شناسایی کند. در ابتدا، با استفاده از یک آنالیزور لغوی ، ابزار به اسکن فایل کد منبع می‌پردازد و برای هر خط از کد ، یک خط تبدیل شده ایجاد می‌شود. کد منبع به پارامترها تبدیل می‌شود؛ و این فرایند شامل تبدیل شناساگر ها و ثابت ها به سمبل (علامت) مشابه p ، و فهرستی از کاندیداهای پارامتر است. برای مثال خط $x=fun(y)+۳*x$ به $P=P(P)+P*P$ تبدیل می‌شود و فهرستی شامل x، fun، y، ۳، و x ایجاد می‌شود. سپس آنالیزور لغوی یک عدد صحیح را ایجاد می‌کند که هر خط کد تبدیل شده را نشان می‌دهد. با داشتن یک طول آستانه، Dup به یافتن انطباق های p کد منبع میپردازد. شباهت به واسطه محاسبه انطباق های p میان دو فایل تعیین می‌شود. Dup در صد شباهت میان دو فایل را به صورت خروجی ارائه می‌کند ، یک پروفایل که نشان دهنده خطوط کد انطباق یافته p است و یک نمودار که نشان دهنده موقعیت کد انطباقی میباشد. با این حال، بنا به گفته گیتچل و تران ، این را در صورت درون گذاری جملات کاذب و یا در صورت بازآرایی بلوک کوچکی از جملات ، نمی تواند فایل‌ها را شناسایی کند. همچنین الگوریتم های انطباق پارامتری دیگری نیز در مقالات وجود دارند.[۲۰]

۳-۴-۴- الگوریتم های مقایسهٔ درخت‌های تجزیه

الگوریتم های مقایسه درخت تجزیه[۳۲] [۲۱]می‌توانند روش‌های مقایسه‌ای را به کار گیرند که ساختار فایل‌ها را با هم مقایسه می‌کنند. برنامه کاربردی کمکی SIM هر فایل را به صورت یک درخت تجزیه نمایش می‌دهد و سپس با استفاده از نمایش رشته‌ای درخت‌های تجزیه ، زوج‌های فایل‌ها را از نظر مشابهت با هم مقایسه می‌کند. فرایند مقایسه با تبدیل هر فایل کد منبع به یک رشته از نشانه‌ها آغاز می‌شود که در آن هر نشانه با یک مجموعه ثابت ثابت از نشانه‌ها جایگزین می‌گردد. پس از نشانه گذاری (نشانه سازی)[۳۳] ، جریان‌های

---

[۳۲] Parse Tree

[۳۳] Tokenization

نشانه‌ای فایل‌ها به بخش‌های مختلف تقسیم می‌شوند و این بخش‌ها هم‌تراز می‌گردند. این روش امکان شناسایی قطعه‌های کد به هم ریخته را فراهم می‌آورد.

SIM از الگوریتم‌های انطباق رشته معمولی برای مقایسه زوج‌های فایل‌های نشان داده شده به صورت درخت‌های تجزیه استفاده می‌کند. همانند اغلب الگوریتم های انطباق رشته‌ای، SIM نیز زیر- دنباله نشانه‌های مشترک ماکسیمال را جستجو می‌کند. درجه شباهت میان زوج‌های فایل‌ها (در دامنه ۰,۰ تا ۱,۰) به صورت خروجی ارائه می‌شود. SIM برای شناسایی شباهت در برنامه‌های C ایجاد شده است. براساس گفته سازندگان SIM، این برنامه کمکی به راحتی قابل انطباق با برنامه‌های نوشته شده به زبان‌هایی به جز C است ؛ و می‌تواند فایل‌های مشابه دربرگیرندۀ اصلاحات و تغییرات رایجی مانند تغییر نام، بازآرایی توابع و جملات، و افزودن/حذف کامنت ها و فضاهای خالی را شناسایی کند.

یکی از کاربردهای اخیر الگوریتم های مقایسه درخت‌های تجزیه چیزی است که در سیستم Brass اجرا می‌شود. کاربرد Brass مستلزم اجرای هر برنامه بصورت یک «جدول ساختاری» است ، که یک نمایش گرافیکی متشکل از نمایش درختی هر فایل است. ریشه درخت تعیین‌کننده هدر (header) فایل است، و گره‌های child تعیین‌کننده جملات و یک جدول علائم (دیکشنری داده ای) که دربرگیرنده اطلاعاتی در مورد متغیرها و ساختارهای داده مورد استفاده در فایل است، تبدیل می‌شود. مقایسه در Brass یک فرایند سه بخشی است. با الگوریتم اول دربرگیرنده مقایسه درخت‌های ساختاری فایل‌ها و الگوریتم سوم دربرگیرنده مقایسه دیکشنری داده‌ها مرتبط با فایل‌ها است. در ابتدا، زوج‌های فایل‌های مشابه با استفاده از الگوریتم مقایسه اول شناسایی می‌شوند. پس از آن، الگوریتم‌های مقایسه دوم و سوم برای زوج‌های فایل‌های شناسایی شده اعمال می‌شوند.

مقایسه درختی مستلزم استفاده از الگوریتم هایی پیچیده‌تر از الگوریتم‌های انطباق رشته‌ای است . با توجه به این مسئله می‌توان فرض کرد که سیستم‌های مبتنی بر سه الگوریتم مقایسه‌ای، کندتر از سیستم‌های مبتنی بر الگوریتم‌های انطباق رشته‌ای هستند. روش فیلترینگ Brass فرایند مقایسه را تسریع می‌کند.

براساس گفته موزگووی ، تحقیقات زیادی در حوزه الگوریتم‌های درخت‌های تجزیه انجام نشده است و مشخص نیست آیا این الگوریتم با عملکردی بهتر از الگوریتم‌های انطباق رشته‌ای از نظر شناسایی زوج‌های فایل‌های مشابه کد منبع دارند یا خیر.

Marble: ازآنجاکه هیچ مقاله انگلیسی در مورد Marble منتشر نشده است در اینجا ویژگی‌های این ابزار را بیشتر توصیف می‌کنیم.

Marble ابزاری است که در سال ۲۰۰۲ در دانشگاه اولترخت طراحی شده .هدف از آن ایجاد یک ابزار ساده با قابلیت پشتیبانی آسان بود که می‌توانست برای شناسایی شباهت‌های مذنون بین کدهای Java مورد استفاده قرار گیرد. با جمع‌آوری همه برنامه‌های ارائه شده برای کارکردهای مختلف در دپارتمان علوم کامپیوتر در اولترخت و به واسطه مقایسه در میان آن‌ها، Marble در نشان دادن سرقت ادبی کارکرد مناسبی را ارائه می‌کرد. به واسطه مقیاس پذیری، بسیار مهم است که این ابزار می‌تواند بین کدهای قدیمی و جدید تمایز قائل شود.[۵]

Marble از یک روش ساختاری برای مقایسه کدها استفاده می‌کند. این کار با تفکیک کد به فایل‌هایی آغاز می‌شود به گونه‌ای که هر فایل تنها دربرگیرنده یک دسته سطح بالا باشد. مرحله بعد، مرحله نرمالسازی است تا جزئیاتی از این فایل‌ها که به‌راحتی توسط دانشجویان قابل تغییر است، حذف شوند: یک آنالیز لغوی انجام میشود که کلمات کلیدی (مانند class، for) و دسته پرکاربرد و نام‌های روش (مانند String، System، toString) را حفظ می‌کند. کامنت ها، فضاهای خالی اضافی، ثابت‌های رشته و تعریف‌های مهم حذف می‌شوند، سایر نشانه‌ها به نشانه آن‌ها "تایپ"، انتزاع می‌گردند. برای مثال، هر عدد مبنای ۱۶ با H و هر کاراکتر لفظی با L جایگزین می‌شوند.

برای هر فایل، Marble دو نسخه نرمال سازی شده را محاسبه می‌کند: یکی که مرتبه فیلدها، روش‌ها و دسته‌های داخلی در آن دقیقاً مشابه فایل اصلی است، و یکی که در آن فیلد ها، روش‌ها و دسته‌های داخلی با یکدیگر گروه بندی می‌شوند و این گروه‌ها مرتب می‌گردند. مرتب سازی به صورت شهودی انجام می‌شود.

برای مثال، روش‌های تعداد از طریق تعداد نشانه‌ها، سپس به واسطه طول کلی رشته نشانه‌ها، و در نهایت به صورت الفبایی، مرتب می‌گردند .

به منظور استخراج دسته‌های داخلی، روش‌ها و فیلد ها از فایل دسته Java بدون الزام به تجزیه، فایل Marble ابتدا به آکولادهای { و } در برنامه به عمق تودرتوی آن‌ها تفسیر می‌شود، و دسته‌ها به واسطه انطباق با آکولادهای عمق تودرتوی راست تفکیک می‌گردند (عمق ۰ متناظر با آکولادهای تعریف دسته و عمق ۱ متناظر با روش‌ها و دسته‌های داخلی هستند). دانستن موقعیت آکولاد باز یک روش نمی‌تواند به طور مستقیم موقعیت آغازین روش به دست دهد، بلکه یافتن آن تنها با اسکن رو به عقب به سمت اولین نقطه-ویرگول یا آکولاد بسته امکان‌پذیر است. دسته‌های داخلی نیز به‌صورت مشابه پردازش می‌شوند.

در هنگام اجرای ابزار می‌توان مقایسه را در میان نسخه‌های نرمال سازی شده مرتب شده یا مرتب نشده (یا هر دو) انجام داد. از آنجا که مرتب سازی شهودی است، تغییرات کمی در روش در یک دسته می‌تواند به طور کامل مرتب سازی روش‌ها را تغییر دهد. موردی مشاهده شده است که در آن یک دانشجو تعدادی از تغییرات را انجام داده بودم و روش‌ها را بازآرایی نکرده بود. به دلیل تغییرات، روش‌های متناظر در نسخهٔ اصلی و نسخه سرقت شده دارای موقعیت‌های بسیار متفاوتی بودند که باعث تأثیرگذاری منفی بر اسکور شد. به همین دلیل، مقایسه نسخه‌های مرتب نشده نیز منطقی است.

۳-۵- مرور کوتاهی بر سرقت ادبی در صفحات وب

اینترنت هم اکنون یک فاکتور کلیدی در زندگی روزمره برای جمع‌آوری اطلاعات و استخراج داده‌ها و مفید از صفحات وب است. [۳]عملکرد و قابلیت اطمینان موتورهای جستجوی وب و دشواری‌های قابل توجهی به واسطه وجود مقادیر بسیار زیادی از داده‌های وب مواجه شده است. حجم بسیار زیاد اسناد کپی برداری شده و شبه کپی برداری شده منجر به ایجاد سربارهای مضاعف در موتورهای جستجو شده است، و کارکرد آن‌ها را شدیداً تحت تأثیر قرار داده است. نیاز به هم گذاری داده‌ها از منابع ناهمگن و مختلف ، به ایجاد صفحات وب شبه کپی برداری شده انجامیده است. داده‌های شبه کپی برداری شده دارای شباهت زیادی با یکدیگر هستند اما در حقیقت «یکسان» نمی‌باشند. اسناد شبه کپی برداری شده در حقیقت صفحات وبی هستند که از نظر محتویات کمی با هم متفاوت‌اند. تفاوت میان این اسناد می‌تواند ناشی از عناصر گنجانده شده در آن‌ها باشد و نه خود محتویات اصلی صفحه. برای مثال، تبلیغات صفحات وب و یا برچسب‌های زمانی مربوط به زمان بروز رسانی یک صفحه وب، هر دو اطلاعاتی هستند که در هنگام جستجوی یک صفحه برای کاربر اهمیتی ندارند، و در نتیجه در هنگام مرور صفحات وب اطلاعات زیادی را به دست نمی‌دهند. وجود صفحات وب شبه کپی برداری شده ناشی از استفاده از مطالب سایت اصلی، سایت انعکاسی، سایت نسخه‌برداری شده، و نمایش‌های گوناگونی از یک آبجکت فیزیکی و اسناد سرقت شده می‌باشد . در بسیاری از شرایط مختلف، دو سندی که دقیقاً قابل تفکیک از یکدیگر نیستند ممکن است حاوی مطالب یکسانی باشند و آن‌ها را باید مطالب شده کپی برداری شده دانست. برای مثال، صفحات وب مربوط به سایت‌های انعکاسی مختلف ممکن است تنها از نظر سربرگ و پانوشت با هم متفاوت باشند. این‌گونه سندها از نظر مضامین سایت با هم فرقی ندارند بلکه در بخش‌های کوچک‌تر مانند تبلیغات و غیره متفاوت می‌باشند.

تشخیص اسناد شبه کپی برداری شده یک مسئله تحقیقاتی بسیار مهم در سال‌های اخیر بوده است. برنامه‌های کاربردی متعددی وجود دارند که می‌توانند به شناسایی اسناد شبه کپی برداری شده در حوزه

تشخیص سرقت ادبی، تشخیص اسپم، و نیز در مرور متمرکز وب کمک کنند. در تشخیص سرقت ادبی، بخشی از یک سند که ممکن است یک جمله یا پاراگراف باشد، در سند دیگری گنجانده شده است و می‌توان آن دو سند را کپی یکدیگر دانست. پیام‌های اسپم متعلق به یک برنامه تبلیغاتی ممکن است بسیار متفاوت به نظر برسد زیرا اسپمرها اغلب باید پیام‌های جدیدی را منتقل کنند و عبارات یا پاراگرافهای نامرتبطی را برای عبور از فیلتر ها استفاده نمایند. با این حال این اسپم‌ها به راحتی از طریق روش‌های تشخیص اسناد شبه کپی برداری شده قابل کشف هستند. تعیین صفحات وب شما به کپی برداری شده به مرور متمرکز وب کمک می‌کند و از کیفیت و تنوع بالاتر نتایج جستجو اطمینان حاصل می‌نماید.

۳-۵-۱- تحقیقات انجام‌شده

روش مربوط به ارزیابی درجه شباهت میان زوج‌های اسناد با عنوان شینگلینگ شناخته می‌شود. برودر و همکارانش روشی را برای انجام این کار پیشنهاد کردند که در آن همه دنباله های کلمات مجاور استخراج می‌شوند. اگر دو سند دربرگیرنده مجموعه‌های شینگل یکسان باشند، آن‌ها را مشابه تلقی می‌کنند و اگر شینگل ها با هم تداخل یا تلاقی کنند، آنها را مشابه های همسان می‌نامند. بااین‌حال محققین دریافتند که این روش برای اسناد کوچک عملکرد مناسبی ندارد. [۳] فترلی و همکارانش از ۵-گرم بعنوان یک شینگل[۳۴] استفاده کردند و نمونه‌ای از ۸۴ شینگل را برای هر سند مورد استفاده قرار دادند. [۳] سپس این ۸۴ شینگل در شش فرا-شینگل قرار می‌گیرند. اسنادی که دارای دو فرا-شینگل مشترک باشند، اسناد شبه کپی برداری شده تلقی می‌شوند. برودر روش مؤثری را برای تعیین شباهت ساختاری فایل‌ها ارائه کرده است و آن را برای هر سند در شبکه جهانی وب (www) مورد استفاده قرار داده است. او با استفاده از این مکانیسم، دسته بندی همه اسناد مشابه از نظر ساختاری را انجام داده است.

روش دیگری برای تشخیص و حذف صفحات وب شبه کپی برداری شده، که با عنوان اطلاعات متنی مبتنی بر اولویت بندی شناخته می‌شود، در مقاله لینگ، هگزین و گواردیان پیشنهاد شده است. با این روش،

---

[۳۴] Shingle

الگوریتمی برای استخراج اطلاعات متنی از صفحات وب توسط درخت DOM و یک الگوریتم اولویت محور برای تشخیص اطلاعات متنی شبه کپی برداری شده اجرا می‌شوند تا نویز صفحات وب را کاهش دهند و در نتیجه کارآمدی و بازدهی تشخیص اطلاعات متنی شبه کپی برداری شده را ارتقا بخشند. [۳]

نارایانا و همکارانش ، روشی را برای تشخیص صفحات وب شبه کپی برداری شده در مرورگری وب پیشنهاد کردند. صفحات وب شبه کپی برداری شده با ذخیره‌سازی صفحات مرورگری شده در منابعی شناسایی می‌شوند. کلمات کلیدی از صفحات مرورگری شده استخراج شوند و بر مبنای این کلمات کلیدی، نمره شباهت میان دو صفحه وب محاسبه می‌شود. در صورتی که نمره شباهت به یک حد آستانه برسد، اسناد مورد نظر مشابه تلقی می‌شوند. با این حال، این محققین از اطلاعات ساختاری صفحات وب که برای دریافت اطلاعات معنایی پیرامون یک صفحه اهمیت دارند، استفاده نکردند. [۳]

متیو ، شین داس و ویجایارانگاوان یک ایده جدید را برای تشخیص صفحات وب شبه کپی برداری شده پیشنهاد کردند که باز هم از الگوریتم مرحله درختی استفاده می‌کند و ارزیابی شباهت در این الگوریتم بر مبنای تجزیه مقدار منفرد (SVM) با استفاده از یک زاویه آستانه θ می‌باشد. با این حال SVM نیازمند عملکردهای ریاضی پیچیده‌تری در ماتریس TDW به همراه تبدیل آستانه Jaccard، t، به زاویه آستانه θ می‌باشد. این مسئله باعث افزایش پیچیدگی‌های الگوریتم و مشکلات عملی در اندازه‌گیری زاویه می‌شود. متیو ، شین داس و ویجایارانگاوان روش جدیدی با نام MWO را برای ارزیابی شباهت مورد استفاده قرار دادند که به طور مستقیم روی آستانه Jaccard، t، کار می‌کند و پیچیدگی الگوریتم را کاهش می‌دهد. [۳]

آن‌ها در این روش، ویژگی‌های معنایی، محتویات و ساختار یک صفحه وب را هم‌زمان بررسی می‌کنند. الگوی وزن گذاری پیشنهادی توسط خود آن‌ها در مقاله‌ای دیگر برای ایجاد یک ماتریس وزن گذاری سند TDW به کار گرفته می‌شود که نقش مهمی را در الگوریتم پیشنهاد دارد. یک الگوریتم سه مرحله‌ای مورد استفاده قرار گرفته است که یک رکورد ورودی را به همراه یک مقدار آستانه دریافت کند و یک مجموعه بهینه از شبه کپی برداری ها را به دست می‌دهد. در مرحله اول، که مرحله نمایش است، همه پیش پردازش

ها انجام می‌شوند و یک الگوی وزنی مورد استفاده قرار می‌گیرد. سپس مرتب سازی به طور کلی به همراه روش‌های نرمال سازی استاندارد برای ایجاد یک ماتریس TDW انجام می‌شود. در مرحله دوم، که مرحلهٔ فیلترینگ است، دو مکانیسم معروف فیلترینگ با عناوین فیلترینگ پیشوند و فیلترینگ محل قرارگیری برای کاهش اندازه مجموعه رقابتی رکورد به کار گرفته می‌شوند و در نتیجه تعداد مقایسه ها کاهش پیدا می‌کند. در مرحله سوم، که مرحله ارزیابی است، بررسی شباهت‌ها با ارائه روش جدیدی به نام تداخل وزنی مینیمم (MWO) بر مبنای یک مقدار آستانه به انجام می‌رسد، و در نهایت یک عدد بهینه از رکوردهای شبه کپی برداری شده حاصل می‌شود.

خلاصه فصل سوم

سرقت ادبی در فایل‌های کد منبع زمانی اتفاق می‌افتد که کد منبع بدون اشاره به مالک و نویسندهٔ اصلی آن کپی برداری و ویرایش شود. سرقت ادبی کد منبع را میتوان بصورت انتقال بخشی از کد منبع نوشته شده توسط فردی دیگر به کد شخصی (بدون اشاره به اینکه کدام بخش‌ها از محقق دیگری کپی شده اند) تعریف کرد.سرقت ادبی معمولاً در محیط‌های آکادمیک اتفاق می‌افتد. دانشجویان به صورت عمدی یا غیرعمدی بخشی از منبع را بدون ذکر نام، در اثر خود می‌گنجانند. تشخیص سرقت ادبی به صورت دستی در مجموعه‌ای از صدها اثر غیرممکن و غیراقتصادی است. بدین ترتیب، ابزارهای نرم‌افزاری برای کمک به مدرسین در تشخیص سرقت ادبی به‌وجود آمده‌اند.

ابزارهای مختلفی برای شناسایی سرقت ادبی وجود دارند که می‌توانند بر مبنای الگوریتم، تقسیم بندی شوند:

سیستم‌های مبتنی بر اثر انگشت (فینگرپرینت) : فینگرپرینت ها در سیستم‌های «شمارش مشخصه» به‌کارگرفته می‌شدند. اولین سیستم شناخته شده تشخیص سرقت ادبی یک برنامه شمارش مشخصه بود برای شناسایی آثار یکسان یا نزدیک به هم مربوط به دانشجویان پدید آمد. بعدها برای شناسایی سرقت ادبی از شمارش عملگرها و عملوندها (operator and operand) استفاده کرد. شاخص‌های به کار گرفته شده : تعداد عملگرهای منحصربه‌فرد ، تعداد عملوندهای منحصربه‌فرد ، تعداد کلی وقوع عملگر ها ، تعداد کلی وقوع عملوندها . روشهای فینگرپرینت نتایج مناسبی را بدست می دهند اما در صورتی که بخش سرقت شده تغییر کرده باشد (مثلاً کلمات جابجا شده یا جایگزین شده باشند) با مشکل مواجه میگردد.

الگوریتم های انطباق رشته : جدید ترین سیستم‌های تشخیص سرقت ادبی متکی هستند بر مقایسه ساختار برنامه‌ها. این سیستم‌ها از شاخص‌های شمارش مشخصه استفاده می‌کنند اما همچنین ساختار برنامه را نیز به منظور بهبود فرآیند تشخیص سرقت ادبی، مقایسه می‌نمایند. سیستم‌های مبتنی بر انطباق رشته از الگوریتم های مقایسه‌ای استفاده می‌کنند که پیچیده‌تر از الگوریتم های شمارش مشخصه هستند. اغلب

الگوریتم های انطباق رشته واسطه تبدیل الگوریتم ها به نشانه‌ها عمل می‌کنند و سپس از یک الگوریتم پیچیده جستجو برای شناسایی زیر-رشته‌های مشترک متنی بین و برنامهٔ استفاده می‌نمایند .

الگوریتم های انطباقی پارامتری سازی شده : الگوریتم های انطباق پارامتری سازی شده همانند روش‌های متداول انطباق نشانه هستند، اما نشانه ساز پیشرفته‌تری دارند. اساساً این الگوریتم های انطباق پارامتری سازی شده با تبدیل فایل‌ها به نشانه‌ها کار می‌کنند.

الگوریتم های مقایسهٔ درخت‌های تجزیه : الگوریتم های مقایسه درخت تجزیه می‌توانند روش‌های مقایسه‌ای را به کار گیرند که ساختار فایل‌ها را با هم مقایسه می‌کنند. درخت های تجزیه میتوانند امکان مقایسه سطح بالا را فراهم کنند. مثلاً میتوانند جملات شرطی را نرمال سازی کنندو ساختارهای هم-ارز را بشناسند. یکی از مزایای آنها میتواند تعیین ساختار یک برنامه باشد. با این وجود، ساخت زیر-درخت ها کار آسانی نیست.

روش	معایب	مزایا
فینگرپرینت	در صورتی که بخش سرقت شده تغییر کرده باشد (مثلاً کلمات جابجا شده یا جایگزین شده باشند) با مشکل مواجه میگردد.	ارائه نتایج مناسبی
الگوریتمهای انطباق رشته ای	ممکن است بواسطه تغییر نام شناساگرها ی اغتشاش و سردرگمی در آنها رخ دهد	• روش های مبتنی بر رشته ها قادر هستند مخدوش سازی های پیچیده بخش سرقت شده (مثل بازآرایی، ادغام، تفکیک جملات، و تغییر مفهومی کلمات) و نیز کپی برداری های مستقیم را تشخیص دهند • این روشها عملکرد سریعی دارند
JPlag	مستلزم تجزیه مجموعه های داده ای است و اگر یک برنامه نتواند تجزیه شود، آن را از مجموعه داده ای حذف میکند.	بعنوان یک سرویس وب در دسترس است و یک اینترفیس کاربری بسیار توانمند برای فهم نتایج دارد. این ابزار از نظر منابع پر بازده عمل میکند و برای سرقت های گسترده میتواند بشکلی مناسب مورد استفاده قرار گیرد.
درخت های تجزیه	نگهداری این درخت مستلزم حافظه زیاد هنگام اجرا این روش است	درخت های تجزیه میتوانند امکان مقایسه سطح بالا را فراهم کنند. مثلاً میتوانند جملات شرطی را نرمال سازی کنندو ساختارهای هم-ارز را بشناسند. یکی از مزایای آنها میتواند تعیین ساختار یک برنامه باشد. با این وجود، ساخت زیر-درخت ها کار آسانی نیست.
SIM	این روش دیگر مورد پشتیبانی فعال سازندگان نیست. درجه یا شاخص شباهت مطرح شده بین ۰ تا ۱ ، بدلیل ناتوانی روش برنامه نویسی دینامیک برای کنترل اجزای کد ترانهاده، با مشکل مواجه میشود.	برنامه نویسی دقیق این روش اطمینان حاصل میکند که زمان محاسباتی در محدوده مناسب قرار می گیرد.

جدول ۶ : مزایا و معایب روش های تشخیص سرقت ادبی در کد نویسی

فصل چهارم : نتیجه گیری و چالش ها

۴-۱- نتیجه گیری

سرقت ادبی عبارت است از سوء استفاده، انتشار، اقتباس یا جایگزینی آثار ، عقاید، و تفکرات دیگران بنام خود و عدم رعایت مالکیت معنوی آن‌ها. برای جلوگیری از فرایند سرقت ادبی، تنها یادآوری این مسئله به دانشجویان که سرقت ادبی کار درستی نیست، کافی نخواهد بود. فرایند شناسایی سرقت ادبی کاری است که باید برای حداقل سازی این عمل نادرست انجام شود.

با افزایش استفاده از اسناد دیجیتال و اینترنت، سرقت ادبی نیز در حال رشد است. از این دیدگاه، ما همچنین شاهد رشد روش‌های تشخیص سرقت ادبی بوده‌ایم و پیشرفت‌های زیادی را در زمینه تشخیص اتوماتیک سرقت ادبی مشاهده کرده‌ایم. روش‌های گوناگونی برای تشخیص سرقت ادبی در حیطه‌های تک زبانی و چند زبانی، روش‌های مبتنی بر گراف، روش‌های معنایی، روش‌های برداری، روش‌های آماری، و غیره ارائه شده‌اند.

۴-۲- چالش ها

ابزارهای کنونی مقابله با سرقت زدگی برای سازمان‌های آموزشی، ناشران و افراد آکادمیک عموماً می‌توانند سرقت ادبی کلمه به کلمه را تشخیص دهند. در حقیقت سرقت ایده‌ها بسیار رایج تر از انواع دیگر سرقت‌های عددی است زیرا افراد آکادمیک زمان کافی را برای کشف ایده‌های جدید ندارند و ناشران ممکن است ابزارهای کافی را برای تشخیص سرقت ادبی انجام شده ندارند. با پیچیده‌تر شدن انواع سرقت‌های ادبی، به نظر می‌رسد که لازم باشد ابزارهای موجود در جهت کشف سرقت‌های ادبی، به حوزه سرقت‌های ادبی معنایی، ساختاری و محتوایی بیشتر بپردازند و یا آنکه ابزارهای جدیدی در این راستا پدید آیند. باتوجه به بررسی ها ی انجام شده در اکثر تحقیقات در این وادی دو فاکتور سرعت و دقت دارای اهمیت به سزای است. اما این روش ها اکثرا به منظور در تشخیص جمله ها دچار مشکل هستند. همچنین یکی دیگر از چالش های اصلی در زمینه سرقت ادبی این است که در اثر الگوریتم های موجود برای نمایش متن از مدل فضای بردار و n_gram استفاده میکنند در حالی این روشها برخی از اطلاعات متن را نادیده میگیرند.

چالش دیگر، روش ها موجود به منظور بررسی متون کوتاه می باشند در صورتی که برای پیاده سازی قانون کپی رایت نیاز به بررسی متون طولانی نیز وجود دارد. در صورتی که با افرایش حجم داده ورودی به روش های فوق یا امکان اجرایی شدن الگوریتم از بین می رود یا زمان اجرای آن غیر قابل قبول خواهد شد.

عدم وجود جامعه استاندار به منظور انجام آزمایشات با کیفیت جهت مفایسه و ارزیانی روش های مختلف در محیطی یکسان را می توان یکی دیگر از چالش های روبرو نامید.

منابع

.۱ Cosma, G. and M. Joy, *An approach to source-code plagiarism detection and investigation using latent semantic analysis.* Computers, IEEE Transactions on, : p. .۳۹۴-۳۷۹, ۲۰۱۲. ۶۱(۳)

.۲ Mozgovoy, M., T. Kakkonen, and G. Cosma, *Automatic student plagiarism detection: future perspectives.* Journal of Educational Computing Research, p. .۵۳۱-۵۱۱ ۲۰۱۰. ۴۳(۴)

.۳ Das, S.N., M. Mathew, and P.K. Vijayaraghavan. *An Efficient Approach for Finding Near Duplicate Web pages using Minimum Weight Overlapping Method*. in *Information Technology: New Generations (ITNG), ۲۰۱۲Ninth International Conference on*. ۲۰۱۲. IEEE.

.۴ Maurer, H.A., F. Kappe, and B. Zaka, *Plagiarism-A Survey.* J. UCS, : p. .۱۰۸۴-۱۰۵۰,۲۰۰۶. ۱۲(۸)

.۵ Hage, J., P. Rademaker, and N. Vugt, *A comparison of plagiarism detection tools.* Utrecht University. Utrecht, The Netherlands: p. .۲۸, ۲۰۱۰

.۶ Arabyarmohamady, S., H. Moradi, and M. Asadpour. *A coding style-based plagiarism detection*. in *Interactive Mobile and Computer Aided Learning (IMCL), ۲۰۱۲International Conference on*. ۲۰۱۲. IEEE.

.۷ Osman, A.H., et al., *An improved plagiarism detection scheme based on semantic role labeling.* Applied Soft Computing, p. .۱۵۰۲-۱۴۹۳, ۲۰۱۲. ۱۲(۵):

.۸ Røkenes, H.D., *Graph-based Natural Language Processing: Graph edit distance applied to the task of detecting plagiarism.* .۲۰۱۲

.۹ El Tahir Ali, A., et al. *Using Kohonen Maps and Singular Value Decomposition for Plagiarism Detection*. in *Computational Intelligence, Communication Systems and Networks (CICSyN), ۲۰۱۱Third International Conference on*. ۲۰۱۱. IEEE.

.۱۰ Ali, A.M.E.T., H.M.D. Abdulla, and V. Snásel. *Overview and Comparison of Plagiarism Detection Tools*. in *DATESO*. .۲۰۱۱

.۱۱ Stamatatos, E. *Plagiarism detection based on structural information*. in *Proceedings of the ۲۰th ACM international conference on Information and knowledge management*. ۲۰۱۱. ACM.

.۱۲ Ceglarek, D. and K. Haniewicz. *Fast plagiarism detection by sentence hashing*. in *Artificial Intelligence and Soft Computing*. ۲۰۱۲. Springer.

.۱۳ Sánchez-Vega, F., et al., *Determining and characterizing the reused text for plagiarism detection.* Expert Systems With Applications, p. ۱۸۱۳–۱۸۰۴, ۲۰۱۳. ۴۰(۵):.

.۱۴ Fang, J. and Y. Zhang, *An Improved Plagiarism Detection Method: Model and Sample*, in *Emerging Technologies for Information Systems, Computing, and Management*. Springer, . p. ۹۶۰–۹۵۳, .۲۰۱۳

.۱۵ BarróN-CedeñO, A., P. Gupta, and P. Rosso, *Methods for cross-language plagiarism detection.* Knowledge-Based Systems, . ۵۰: p. .۲۱۷–۲۱۱, ۲۰۱۳

.۱۶ Potthast, M., et al., *Cross-language plagiarism detection.* Language resources and evaluation, . ۴۵(۱): p. .۶۲–۴۵, ۲۰۱۱

.۱۷ Bao, J.P., C. Lyon, and P.C. Lane, *Copy detection in Chinese documents using Ferret.* Language resources and evaluation, . ۴۰(۴–۳): p. ۳۶۵–۳۵۷, .۲۰۰۶

.۱۸ Khan, M.A., et al. *Copy detection in urdu language documents using n-grams model*. in *Computer Networks and Information Technology (ICCNIT), ۲۰۱۱International Conference on*. ۲۰۱۱. IEEE.

.۱۹ Tresnawati, D. and R. Syaichu, *Plagiarism Detection System Design for Programming Assignment in Virtual Classroom based on Moodle.* Procedia-Social and Behavioral Sciences, . ۶۷: p .۱۲۲–۱۱۴, ۲۰۱۲

.۲۰ Roy, C.K. and J.R. Cordy, *A survey on software clone detection research*, , Citeseer, .۲۰۰۷

.۲۱ Ellis, M.G. and C.W. Anderson, *Plagiarism Detection in Computer Code.* Mar, . ۲۳: p. .۱۰–۱

, ۲۰۰۵

واژه نامه فارسی به انگلیسی

انگلیسی	فارسی
Cluster Based Methods	روش‌های خوشه بندی
Comment	کامنت ها
Comparable Corpus	مجموعه‌های داده قابل مقایسه
Cosine Similarity Measure	معیار شباهت کسینوسی
Cross-Language Methods	روش‌های چند زبانی
Decomposition Singular Value	روش تجزیه مقدار منفرد
Fingerprints	اثر انگشت ها
Grammer-based methods	روش‌های مبتنی بر گرامر
Lexicon	شباهت‌های لغات
Loop Building	حلقه-ساز
Operand	عملوندها
Operator	عملگرهای
Parallel Corpus	مجموعه‌های داده ای موازی
Parameterized Matching Algorithm	الگوریتم های انطباقی پارامتری سازی شده
Parse Tree	درخت تجزیه
Post-Processing	فراپردازش
Reused Text	متن تکراری
Self Organization Map	نقشه های خود-آرایی
Semantic Methods	روش‌های معنایی (صرفی)
Semantic methods	روش معنایی
Semantic Role Labeling	برچسب زنی نقش معنایی
Sentence Hashing	درهم سازی جملات
Source Code	کد منبع
String Matching	انطباق رشته
String Similarity	شباهت رشته‌ای
Structure-Based Methods	روش‌های مبتنی بر ساختار
Syntactic Models	مدلهای نحوی
Thesaurus	فرهنگ لغت
tile	مجموعه معنی دار
Tokenization	نشانه گذاری
Unified Distance Matrix	ماتریس فاصله یکپارچه