مجله محاسبات نرم، جلد ۱۱، شماره ۱ (بهار و تابستان ۱۴۰۱)، صفحه ۶۰ـ۷۱
https://doi.org/10.22052/scj.2022.243384.1034

مجله محاسبات نرم
SOFT COMPUTING JOURNAL
تارنمای مجله: scj.kashanu.ac.ir

دانشگاه کاشان
University of Kashan

تشخیص سرقت ادبی در متون علمی مبتنی بر بلوک‌بندی متن و معیار مشابهت کسینوسی✧

نگار مجمع[*1]، استادیار، سارا باشتین[1]، دانشجوی کارشناسی ارشد
[1] گروه مهندسی کامپیوتر، موسسه آموزش عالی نقش جهان، اصفهان، ایران.

اطلاعات مقاله

تاریخچه مقاله:
دریافت ۰۶ دی ماه ۱۴۰۰
پذیرش ۰۴ آذر ماه ۱۴۰۱

کلمات کلیدی:
سرقت علمی ادبی
تشخیص اصالت متون علمی
فاصله کسینوسی
بلوک‌بندی متن

چکیده

در دهه اخیر با گسترش دسترسی به شبکه جهانی اینترنت، سرعت و سهولت در دسترسی به ایده‌ها، مستندات، مقالات، دست نوشته‌ها و داده‌های جمع‌آوری شده توسط دیگران افزایش یافته است. این موضوع باعث شده است که تبادل اطلاعات و افکار بین محققین و تولیدکنندگان علوم آسان‌تر شود، اما در مقابل باعث آسان شدن اعمال رونوشت غیرمجاز، خلاصه نویسی بدون ذکر منبع و در کل سرقت متون ادبی شده است. از آنجایی که دانشگاه‌ها و مراکز آموزشی، منابع علمی و پژوهشی را با سهولت در دسترس اغلب کاربران قرار می‌دهند، تشخیص میزان اصالت متون علمی در این مراکز مهم‌تر و بالطبع آن از حساسیت بیشتری برخوردار است. در این پژوهش روشی ارائه شده تا با استفاده از بلاک‌بندی قطعات اسناد، مقایسه بین قطعات مرتبط انجام شود. در روش پیشنهادی پس از دسته‌بندی اسناد به دو دسته اسناد اصلی و اسناد مشکوک، پیش‌پردازشی با هدف حذف ایست واژه‌ها و جمله‌بندی جدید صورت پذیرفته است. سپس اسناد قطعه‌بندی شده و با استفاده از شباهت کسینوسی، میزان شباهت متون با یکدیگر تعیین شده است. روش پیشنهادی در آزمون ۵۰ سند موجود در مجموعه داده‌ها، دقت ۹۴ درصدی را کسب کرده که به نسبت به یکی از روش‌های مشابه بهبود ۲ درصدی داشته است.

۱. مقدمه

سرقت ادبی، تلقی ایده، نظرات و افکار دیگران به نام خود و یا استفاده از آنها بدون ارجاع به منبع اصلی است [۱]. افراد با جابه‌جایی کلمات، جملات و یا عبارات و جایگزینی معانی آنها و همچنین دوباره‌نویسی[1] متن، سعی در پنهان نمودن موارد رونوشت شده از کار دیگران در کار خود را دارند. امروزه، دسترسی آسان به منابع اطلاعاتی، از طریق وب یا شبکه‌های محلی دیگر، موجب شده است تا دانشجویان و دانش‌آموزان به رونوشت غیرمجاز از ایده‌ها و اطلاعات دیگران ترغیب شوند. اغلب موارد رونوشت از آثار ادبی در دانشگاه‌ها است که اسناد به طور معمول به صورت مقاله و گزارش یافت می‌شوند. با این حال، رونوشت از آثار دیگران می‌تواند به صورت نسبی در هر زمینه‌ای شامل مقالات علمی، طراحی‌های هنری و کدهای منبع صورت گیرد [۲]. سرقت علمی می‌تواند در چندین شکل ظاهر شود، استفاده از متن، یک رابطه ریاضی، شکل یا نمودار و غیره بدون ارجاع به صاحب اثر نمونه‌های بارز سرقت هستند. از میان انواع نمونه‌های سرقت ادبی، استفاده از قسمت یا کل یک سند متنی بیشترین میزان رخداد را به خود اختصاص داده است.

✧ نوع مقاله: پژوهشی
* نویسنده مسئول
پست(های) الکترونیک: majma@naghshejahan.ac.ir (مجمع)
sbashtin4@gmail.com (باشتین)

[1] Rewriting

نحوه ارجاع به مقاله: مجمع، نگار، باشتین، سارا، «تشخیص سرقت ادبی در متون علمی مبتنی بر بلوک‌بندی متن و معیار مشابهت کسینوسی»، مجله محاسبات نرم، جلد ۱۱، شماره ۱، ص۶۰ـ۷۱، بهار و تابستان ۱۴۰۱.

هدف متن‌کاوی، کشف اطلاعات ناشناخته و نانوشته است. در حقیقت متن‌کاوی فرآیند استخراج اطلاعات مفید و دانش از متن غیرساخت‌یافته است. متن‌کاوی یک زمینه تحقیقاتی جدید میان رشته‌ای است که به حوزه‌هایی چون بازیابی اطلاعات، داده‌کاوی، یادگیری ماشین، آمار و زبان‌شناسی محاسباتی مربوط است. در واقع، متن‌کاوی نوعی داده‌کاوی است، با این تفاوت که ابزارهای داده‌کاوی برای کاربر روی داده‌های ساخت‌یافته درون پایگاه داده طراحی می‌شود، اما متن‌کاوی با مجموعه‌ای از داده‌های غیرساخت‌یافته و یا نیمه‌ساخت‌یافته مانند رایانامه‌ها، سندهای متنی و فایل‌های HTML کار می‌کند [۳]. تا به امروز بیشتر پژوهش‌ها و تلاش‌ها، بر روی داده کاوی با استفاده از داده‌های ساخت‌یافته متمرکز شده است. یکی از شاخه‌های بسیار جذاب در متن‌کاوی محاسبه خودکار شباهت متون است. با توجه به این که نتایج حاصل از اغلب پژوهش‌های انجام شده در حال حاضر در قالب متن ارائه می‌شود، محاسبه شباهت متون با هدف تشخیص اصالت متن و صیانت از مالکیت مادی و معنوی پژوهش‌های منتشر شده، یکی از الزامات پژوهشی و ساختاری به شمار می‌رود. بر همین اساس داده‌های موجود در اسناد اعم از بخش متن و تصویر استخراج شده و میزان شباهت آنها با اسناد دیگر سنجیده می‌شود تا اصالت سند تشخیص داده شود [۴]. به صورت کلی روش‌های شناسایی سرقت ادبی در متون به دو دسته تقسیم می‌شود: دسته اول که دارای قدمت بیشتری است، مبتنی بر متن و دسته‌ی دوم مبتنی بر محتوا است. تا کنون روش‌های متعددی برای شناسایی سرقت ادبی با هر دو روش ارائه شده است، اما روش‌هایی که بیشترین محبوبیت را در میان پژوهش‌ها به خود اختصاص داده‌اند، روش‌های مبتنی بر ترکیب این دو روش هستند. این بدین معناست که سرقت ادبی نه تنها از متن، بلکه باید از محتوای اثر نیز تحقیق شود. با بررسی ادبیات و پیشینه تحقیق محقق گردید که مشکل اول در تشخیص اصالت متون ادبی، یا به تعبیر دیگر کشف سرقت ادبی که اکثر روش‌های موجود از آن رنج می‌برند، تعداد زیاد اسناد موجود به عنوان اسناد مورد مقایسه است [۲]. این بدین معنی است که زمانی اصالت یک متن قابل اثبات است که متن مورد نظر با تمام متون ثبت شده در پوشه‌های موجود در دنیا مقایسه شده و میزان مشابهت آن سند تعیین گردد. بنابراین این مشکل موجب می‌گردد که تعداد مقایسات و به طبع آن زمان پردازش به حد چشمگیری افزایش یابد. اما با به‌کارگیری روش‌های مقایسه بخش‌هایی از متن به جای کل متن، می‌توان از زمان پردازش کاست. به این صورت که ابتدا قسمتی از سند با اسناد موجود مقایسه شده و در صورت وجود مشابهت با یک حد آستانه مشخص، اسناد مورد نظر برای محاسبه میزان مشابهت کاندید می‌شوند. این امر با وجود بهبود زمان پاسخ‌دهی در کاهش دقت شناسایی سرقت ادبی اصلی‌ترین نقش را ایفا می‌کند.

این پژوهش برای حل این مشکلات در مرحله پیش‌پردازش از راه‌کار بلاک‌بندی متن و مقایسه بلاک‌های اسناد استفاده کرده است. پس از استخراج بلاک‌های متنی، فرآیند محاسبه شباهت کسینوسی بلاک‌های سند مشکوک با بلاک‌های اسناد بازیابی شده (کاندید) انجام می‌شود. سپس مقادیر شباهت کسینوسی بلاک‌ها جهت تعیین شباهت کلی اسناد با سند مشکوک محاسبه شده و در پایان با مرتب‌سازی اسناد مخزن بر اساس شباهت کل میزان مشابهت و سرقت ادبی سند مشخص می‌گردد.

در ادامه، این مقاله در شش بخش سازماندهی شده است. در بخش دوم مروری بر پیشینه تحقیق انجام شده، در بخش سوم روش پیشنهادی با هدف پاسخ به چالش‌های ذکر شده ارائه شده و در بخش چهارم نتایج حاصل از روش پیشنهادی بیان شده است. در بخش پنجم نتایج حاصله ارزیابی شده و در نهایت، در بخش ششم نتیجه‌گیری و جمع‌بندی این تحقیق ارائه شده است.

۲. پیشینه تحقیق

ساروار و همکاران [۱]، روشی برای رتبه‌بندی اسناد را ارائه دادند که بتواند با کاوش و تجزیه و تحلیل معیارهای شباهت، که در آن هر سند به عنوان یک سری از قطعات منحصر به فرد شناسایی می‌شود، رتبه‌بندی را انجام دهد. در این مقاله، از دو مجموعه داده متفاوت بهره برده شده و برای ارزیابی از میانگین

دقت (MAP[1]) و precision@k برای تجربه کاربر و رفتار اسناد مربوطه که بر حسب فرکانس رتبه‌بندی آنها با مقادیر آستانه متفاوت بازگردانده می‌شود، استفاده شده است. لذا هر سند به صورت جملاتی در قالب $d = \{p_1, p_2, p_3, \dots, p_n\}$، نگهداری شده و تلاش برای تخمین بهتر تشابه از طریق چند تابع مشابهت‌سنجی انجام می‌پذیرد.

Alzahrani و همکاران [۵]، برای تشخیص اصالت متون ادبی سیستم جعبه سیاه[2] را معرفی نمودند. ورودی‌های این سیستم یک متن پرس‌وجو و مجموعه اسنادی است که احتمالا رونوشت از آنها صورت گرفته است.

در تحقیق Potthast و همکاران [۶]، فرآیند بازیابی برای تشخیص اصالت متون ادبی خارجی، در سه مرحله صورت می‌گیرد که ورودی آن یک سند پرس‌وجو و منابع شامل منابع اصلی است. سه عملگر اصلی جعبه سیاه بدین صورت است: در مرحله اول یک لیست کوچک از اسناد کاندید که به سند اصلی شباهت دارند از مجموعه منابع بازیابی می‌شود؛ در مرحله دوم سند پرس‌وجو را با اسناد کاندید مقایسه می‌کنند که این کار با استفاده از واحدهای مقایسه‌ای چون k گرام یا جمله انجام می‌شود؛ در نهایت در مرحله سوم، در میان اسناد اصلی کاندید سندهایی که شباهت کافی با سند پرس‌وجو ندارند، از لیست اسناد کاندید حذف می‌شوند. در این روش بیشتر حروف و یا کلمات با یکدیگر تطبیق داده می‌شوند، به همین خاطر باید زبان نوشتاری متون یکی باشد؛ لذا تنها در تشخیص اصالت متون ادبی تک زبانه استفاده می‌شود. همچنین از این روش نمی‌توان در مواردی که سبک نگارش تغییر یافته است، استفاده نمود زیرا این تغییرات در سطح کپی‌برداری دقیق نمی‌باشند. بنابراین این روش مناسب برای تشخیص اصالت متون ادبی غیرذاتی یا بیرونی مناسب است. در این روش از تطبیق رشته و شبیه‌سازی جملات استفاده می‌شود.

Grozea و همکاران [۷]، در تحقیقات خود برای تشخیص اصالت متون ادبی غیرذاتی شکل لغوی را استفاده کردند. لازم به ذکر است که Alzahrani و همکاران [۵] نیز برای تشخیص اصالت متون ادبی غیرذاتی بر مبنای فازی معنایی-شبیه‌سازی رشته از روش n گرام کلمه برای تشخیص شباهت استفاده کردند. علاوه بر آنها، Basile و همکارانش [۸] در تحقیقات خود تحت عنوان ارائه یک رویکرد سه مرحله‌ای انتخاب، تطبیق و مجذور در شکل لغوی متن از روش n گرام کلمه استفاده نمودند.در ادامه، Elhadi و همکارش [۹]، با استفاده از ساختار گرامری متن در تشخیص متن تکراری از این روش استفاده نمودند. همچنین Koroutchive و همکارش [۱۰]، در تشخیص اصالت متون ادبی با ترجمه بعضی از متون و داده‌ها با منبع مشترک از این روش استفاده کرده‌اند. این روش در رونوشت-های حرفی کاربرد دارد.

به عنوان روشی مبتنی بر معنا، Li و همکاران [۱۱]، بر اساس جملات مشابه روی گره معنایی و مجموعه آماری از نوع رونوشت غیرمجاز هوشمندانه دستکاری متن در پاراگراف استفاده کردند. در این زمینه، Bao در سال ۲۰۰۴ در جملات معنایی و تشخیص اصالت متون ادبی در سطح خلاصه متن را نیز با این روش الگویی ارائه دادند که هنوز به اثبات نرسیده است [۱۲].

Alzahrani [۱۳]، مجموعه فازی بر مبنای توضیحات دوباره اطلاعات و تطبیق اثر انگشت برای تشخیص اصالت در متون عربی در سطح رونوشت غیرمجاز حرفی و در سطح رونوشت غیرمجاز هوشمندانه در پاراگراف را استفاده کردند که این روش جواب داده است.

به عنوان روشی مبتنی بر ساختار، Zhang و همکاران [۱۴]، و همچنین Chow و همکارش [۱۵]، از SOM چندلایه و ساختار داده درختی برای استخراج ساختار اسناد بر مبنای محتوا تحقیقاتی را انجام داده‌اند. البته روش ساختاری توانسته اصالت ادبی نوع رونوشت غیرمجاز حرفی را تشخیص دهد، اما نوع هوشمندانه آن در سطح پارگراف و خلاصه‌سازی و بخش ایده و محتوا، الگوهایی ارائه شده که هنوز اثبات نشده است.

Rakian و همکاران [۱۶]، یک رویکرد جدید فازی برای کشف سرقت ادبی در متون فارسی ارائه دادند. رویکرد پیشنهادی در

[1] Mean Average Precision (MAP)
[2] Black box

این تحقیق بر روی متون paraphrase شده و با هدف شناخت شباهت‌های متن می‌باشد.

تحقیق Gupta و همکاران [۱۷] برخلاف روش‌های دیگر که به طور معمول به کلمات محتوا سند تکیه می‌کنند، با استخراج کلمات اضافی مانند (of, the, and, or) انجام می‌شود. روش پیشنهادی با اضافه کردن مرزهای جمله صحیح به همراه پروفایل‌های n گرام کلمه اضافی پیاده‌سازی شده است.

Sanchez-Perez و همکاران [۱۸]، از یک روش بهینه‌سازی مبتنی بر الگوریتم ژنتیک برای بهبود عملکرد یک مدل تشخیص سرقت ادبی با بهینه‌سازی پارامترهای ورودی آن استفاده نمودند. اجرای الگوریتم ژنتیک آنها مبتنی بر بازنمایی غیردودویی افراد، انتخاب دقیق ژن‌های متن، انتخاب‌های متقاطع یکنواخت و میزان جهش زیاد در میان ژن‌های انتخابی در متن است.

Schubotz و همکاران [۱۹]، برای کشف سرقت ادبی از یک روش مبتنی بر تحقیقات ریاضی استفاده نمودند. در این تحقیق، سرقت ادبی با بررسی یادداشت‌های سرمقاله از zbMATH شناسایی شده است. در حالی که بیشتر مواردی موجود یک کپی ساده از یک نسخه اولیه (مقالات پایه) بوده است، اما در این روش چندین نوع سرقت ادبی نیز شناسایی شده است.

Chang و همکاران [۲۰]، برای تشخیص سرقت ادبی از روش تبدیل بردار کلمه استفاده کرده‌اند که می‌تواند رابطه معنایی بین کلمات مختلف را آشکار کند. این بردارهای کلمه در خوشه‌های مختلف دسته‌بندی می‌شوند. از نظر تشابه، در این مقاله نیز جملات به بلاک‌های دسته‌بندی شده و به جای خوشه‌بندی جملات، به جملات وزن‌دهی صورت گرفته است.

محققان در مقاله [۲۱]، به بررسی روشی جهت استخراج کلمات کلیدی در اسناد وب پرداخته‌اند. روش ارائه شده، تعداد تکرار خام عبارت را در بخش منتخبی از صفحات مشخص می‌نماید.

در مقاله [۲۲]، روشی بر اساس ویژگی‌های سبکی و نوشتاری نویسندگان در زبان فارسی برای تشخیص هویت آنها ارائه شده است. در این مقاله، ویژگی‌های واژگانی و نحوی لغات از متن استخراج شده و مدل بر اساس آنها آموزش داده شده است.

Nagoudi و همکاران [۲۳]، به ارائه روشی مبتنی بر یادگیری ماشین برای تشخیص سرقت ادبی در متون عربی پرداخته‌اند که همانند روش پیشنهادی این مقاله در سطح جمله این بررسی را انجام می‌دهد. اما وزن‌دهی را به کلمات انجام می‌دهد که از نظر حجم پردازش، پردازش سنگینی را خواهد داشت.

Agrawal و همکاران [۲۴]، الگوریتمی برای مقایسه متون مشابه با استفاده از مقادیر هش ارائه کرده‌اند که توجه کمی به ساختار کلمات در متن دارد.

در مقاله [۲۵]، دو روش برای شناسایی سرقت ادبی پیشنهاد شده است که از دو مرحله فیلتر بر اساس روش کیف کلمه[1] در سطح سند و در سطح جمله استفاده می‌کند. در روش اول برای تشخیص شباهت در اسناد و جملات مشکوک، ترکیب روش شبکه‌ای از پیش آموزش داده شده کلمات جاسازی FastText و روش وزن‌دهی TF-IDF برای تشکیل دو ماتریس ساختاری و معنایی و در روش دوم برای تشکیل دو ماتریس، WordNet هستی‌شناسی و وزن‌دهی TF-IDF استفاده شده است. بخش استفاده از وزن‌دهی مورد استفاده در این مقاله، مشابه راه‌کار پیشنهادی ما در این مقاله است.

محققان در مقاله [۲۶]، روشی برای نمایش عددی متن در یک فضای برداری و جاسازی کلمه به همراه معیارهای تشابه به متن پیشنهاد کرده‌اند که بر مبنای الگوریتم‌های یادگیری عمیق مانند شبکه‌های کانولوشنی به تشخیص سرقت ادبی و کشف آن می‌پردازد.

نویسندگان مقاله [۲۷]، به بررسی روش‌های گوناگون بررسی ویژگی‌های خاص متن با روش‌های متفاوت از جمله تحلیل‌های لغوی و نحوی و یا روش‌های ایستا و پویا مانند یادگیری ماشین پرداخته‌اند.

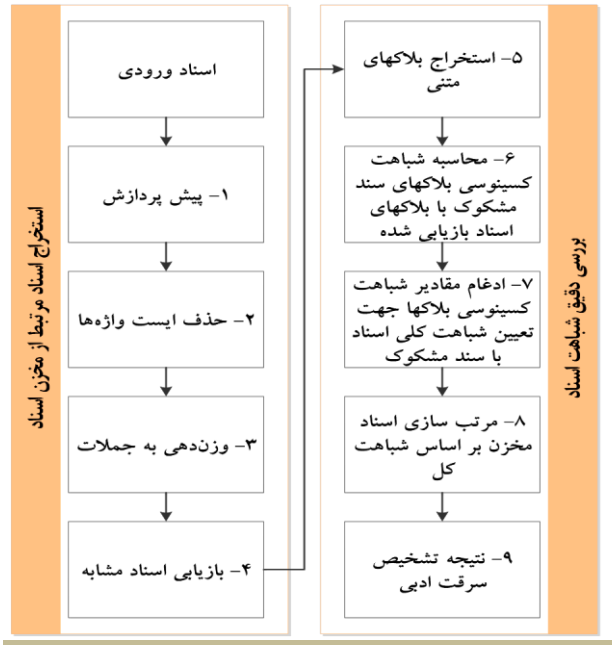
۳. روش پیشنهادی

روش پیشنهادی در این تحقیق شامل ۹ مرحله است، که این ۹ مرحله در شکل (۱) قابل مشاهده است. در ادامه این بخش به تشریح کامل این ۹ مرحله پرداخته خواهد شد.

به طور خلاصه، پس از پیش‌پردازش متن، که خود در سه مرحله

[1] Bag of word

انجام می‌شود، نوبت به بلاک‌بندی می‌رسد. در بلاک‌بندی متن به بلاک‌هایی که هر بلاک یک پاراگراف است، تقسیم می‌شود. سپس نوبت به حذف داده‌های ایستا می‌رسد، این کلمات شامل (a, the, is, if, then) و غیره هستند که در ادامه این کلمات مشخص خواهد شد. سپس کلمات باقی‌مانده جمله‌بندی شده و جملات اولیه برای وزن‌دهی ایجاد می‌گردند. در بخش وزن‌دهی به هر یک از جملات با استفاده از روش TF-IDF یک وزن داده خواهد شد. در ادامه، در بخش بررسی دقیق شباهت اسناد، اسناد بازیابی شده در بخش قبل مجدد بلوک‌بندی می‌شود. سپس شباهت بلوک‌ها به صورت دودویی محاسبه می‌گردد، یعنی هر بلوک از سند مورد مقایسه با تمام بلوک‌های تمام اسناد بازیابی شده مقایسه شده و میزان شباهت بلوک‌ها و در ادامه میزان شباهت اسناد مشخص می‌شود. در پایان اسناد با توجه به میزان شباهت مرتب و اسنادی با بیشترین شباهت بازیابی می‌شود.



شکل (۱): دیاگرام روش پیشنهادی

۳,۱. پیش‌پردازش

در این مرحله، ۳ عملیات انجام می‌شود که در نهایت یک سند به مجموعه‌ای از جملات تبدیل می‌شوند که عبارتند از:

الف. جداسازی اسناد: اسناد اصلی و اسناد مشکوک تفکیک می‌شوند.

ب. استخراج اطلاعات اسناد: بخش‌های مختلف مقاله استخراج می‌شود.

ج. بلوک‌بندی بخش اصلی: بخش اصلی سند به بلوک‌های کوچکتر (پاراگراف) تقسیم می‌شوند.

۳,۱,۱. جداسازی اسناد ورودی

به طور کلی اسناد ورودی به دو دسته، اسناد اصلی و سند مشکوک تقسیم می‌شوند که زبان این اسناد انگلیسی است. اسناد اصلی، شامل اسنادی است که احتمالا رونوشت غیرمجاز از روی این اسناد صورت گرفته است. اسناد مشکوک نیز اسنادی هستند که احتمالا در آنها بخش‌هایی وجود دارد که از روی آنها کپی برداری شده است. این اسناد به عنوان سند پرس‌وجو شناخته می‌شود. در روابط (۱) و (۲)، DS مجموعه اسناد مشکوک و DO مجموعه اسناد اصلی هستند.

$$DO = \{Do_1, Do_2, \dots, Do_n\} \qquad 1 \le i \le n \qquad (۱)$$

$$DS = \{Ds_1, Ds_2, \dots, Ds_m\} \qquad 1 \le i \le m \qquad (۲)$$

۳,۱,۲. استخراج اطلاعات سند

این قسمت با استفاده از روش تبدیل متن به بلوک انجام می‌شود [۵]. این روش با دریافت مقاله به فرمت PDF پردازش را آغاز می‌کند، بدین گونه که ابتدا فایل PDF را خوانده و سپس شروع به استخراج اطلاعات آن می‌کند. اطلاعات مورد نظر به سه نوع تقسیم می‌شوند:

۱. اطلاعات سند: عنوان، نویسنده و تاریخ
۲. اطلاعات صفحه: تعداد صفحات و ابعاد صفحه
۳. خطوط رشته‌ای در صفحه: اندازه، نوع قلم، سبک، رنگ و از این دست اطلاعات

پس از استخراج اطلاعات فوق اسناد با فرمت PDF به قالب متنی (XML) تبدیل می‌شوند زیرا فایل‌ها با فرمت PDF نمی‌توانند شامل اطلاعات پاراگراف‌ها باشند.

۳,۱,۳. بلاک‌بندی

برای شباهت‌سنجی اسناد، در مرحله اول اسناد با فرمت PDF به

فایل‌های XML تبدیل شده و در ادامه این فایل‌ها بلاک‌بندی می‌شود. در بخش بلاک‌بندی هر پاراگراف به یک بلاک تقسیم می‌شود.

۳،۲. شناسایی و حذف کلمات بدون اثر

یکی از راه‌های ساده برای مخفی کردن استفاده غیرقانونی از متون علمی حذف یا اضافه کردن کلمات، فضاهای خالی، علایم و خطوط جدید در متون کپی شده از متن‌های دیگر است. کلمات بدون اثر به کلمات پرتکراری گفته می‌شود که تاثیر کمی در فهم تخصصی متن دارد. با توجه به اینکه این کلمات در اغلب متون تکرار می‌شوند به آنها کلمات کمکی نیز گفته می‌شود. این کلمات شامل انواع ضمیرها، حروف ربطی و اضافه هستند و تاثیری زیادی در ارزش محتوایی اسناد ندارند. تمامی این کلمات از متن حذف شده و سپس کلمات اصلی متن ذخیره می‌شوند. لیست تمامی کلمات بدون اثر در جدول (۱) ارائه شده است.

جدول (۱): لیست کلمات بی‌اثر

1.a	6.at	11.if	16.no	21.such	26.there	31.this
2.an	7.be	12.in	17.not	22.that	27.these	32.to
3.and	8.but	13.into	18.of	23.the	28.they	33.was
4.are	9.by	14.is	19.on	24.their	29.these	34.will
5.as	10.for	15.it	20.or	25.then	30.they	35.with

۳،۳. وزن‌دهی به جملات و شاخص‌گزاری

بعد از حذف کلمات بدون اثر، مجموعه‌ای از کلمات کلیدی اسناد حاصل می‌شود که هر مجموعه از این کلمات در بین دو نقطه جمله نامیده می‌شود. با استناد به این مجموعه‌ها می‌توان وزن هر جمله اصلی را در سند به دست آورد. نسبت‌دهی یک مقدار عددی به هر یک از جملات انتخاب شده را وزن‌دهی به جملات می‌گویند. این کار برای این انجام می‌شود تا تمایز متن از سایر متون نمود بیشتری داشته باشد. مراحل محاسبه وزن توسط روش TF-IDF مطابق روابط (۳) تا (۶) انجام می‌شود.

$$d_j = (w_1, w_2, \dots, w_n) \quad (۳)$$

$$w_{ij} = tf_{ij} \times idf_i = tf_{ij} \times \log_2(\frac{N}{df_i}) \quad (۴)$$

$$tf_{ij} = \frac{f_{ij}}{\max_i\{f_{ij}\}} \quad (۵)$$

$$idf_i = 1 + \log_2(\frac{N}{df_i}) \quad (۶)$$

که در آنها d_j بردار متن j ام، w_{ij} وزن جمله i در متن j ام، f_{ij} تعداد تکرار جمله i در متن j ام، df_i تعداد اسناد از مجموعه آموزشی که جمله i ام در آن حداقل ۱ بار موجود باشد، tf_{ij} تعداد تکرارهای جمله i در متن j ام یا تعداد تکرار هر کلمه، idf_i معکوس متن از تکرار جمله i یا معکوس فرکانس متن و N تعداد کل اسناد است.

با وزن‌دهی به جملات برای هر کدام از آنها یک مقدار عددی جهت شاخص‌گذاری به دست می‌آید، با این کار یک فایل index از سند حاصل می‌شود که اطلاعات کل سند را نمایش خواهد داد. این اطلاعات شامل جملات موجود در هر بلوک، وزن جملات در سند و نهایتا idf سند است.

۳،۴. بازیابی اسناد مشابه

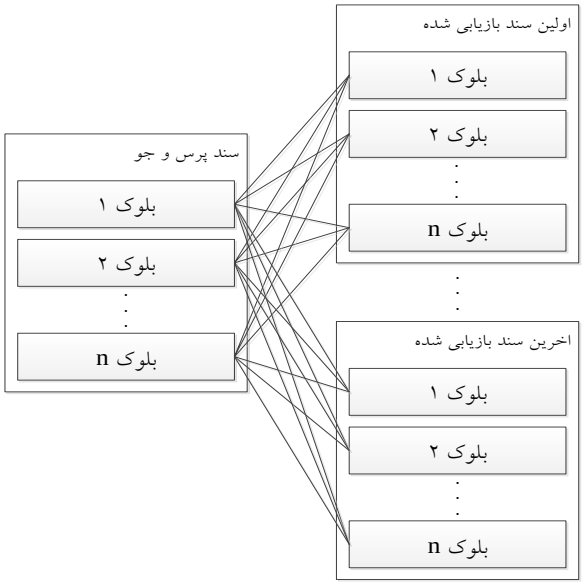
برای بررسی میزان مشابهت دو سند و بازیابی این اسناد از روش شاخص‌گذاری متون استفاده شده است. برای معیار شباهت از مدل فضای برداری (VSM) بهره گرفته شده که با آن، وابستگی و شباهت متون بر اساس بردار مشابهت اسناد با سند پرس‌وجو محاسبه می‌شود. در مدل برداری، متن‌ها به عنوان بردارهایی در فضا نمایش داده می‌شوند. هر بردار می‌تواند به وسیله وزن‌های هر جمله در یک پاراگراف با توجه به بعد فضایی نمایش داده شود. تعداد بعدها مساوی با تعداد جملات استفاده شده هستند. این بخش را می‌توان با روش ماتریسی نمایش داد که ستون‌ها نشان‌دهنده جملات و سطرها نشان‌دهنده اسناد منبع باشند. به عبارتی متن‌ها به شکل بردار در فضای چندبعدی نمایش داده می‌شود که هر محور این فضا متناظر با یک جمله یا عبارت است. مشخصا ابعاد بردارها می‌تواند بسیار زیاد باشد. برای اندازه‌گیری شباهت از روش شباهت کسینوسی استفاده شده که کسینوس زاویه بردار جملات سند پرس‌وجو که با q نمایش داده شده است و مجموعه بردارهای اسناد که با d نمایش داده شده است، از طریق رابطه (۷) محاسبه می‌شود.

$$sim_{cosine}(q,d) = \frac{\sum_{t\in q\cap d}(w_{q,t}\,.\,w_{d,t})}{\sqrt{\sum_{t\in q} w_{q.t}^2} \times \sum_{t\in d} w_{d,t}^2} \qquad (۷)$$

در واقع با این عمل به اسناد یک رتبه‌بندی اولیه تخصیص داده می‌شود و سپس اسناد برتر با استفاده از محاسبه دقیق شباهت بازیابی می‌شوند.

۳,۵. محاسبه دقیق شباهت

در ادامه این بخش، مشابهت هر بلوک از سند پرس‌وجو با بلوک‌های بازیابی شده، مطابق روش بیان شده در بخش قبل محاسبه می‌شود. نمایی از این بخش در شکل (۲) نمایش داده شده است. در این بخش برای هر بلوک از سند پرس‌وجو و هر بلوک از اسناد بازیابی شده یک مقدار به عنوان مقدار مشابهت بازیابی می‌شود. در پایان میزان مشابهت کل سند پرس‌وجو و سند بازیابی شده از مجموع کل مقادیر مشابهت‌های بلوک‌های سند پرس‌وجو و سند بازیابی شده حاصل می‌شود و اسناد با بالاترین میزان مشابهت بازیابی می‌گردند.



شکل (۲): نمایی از بخش محاسبه دقیق شباهت

۳,۶. پیاده‌سازی

روش پیشنهاد شده در تحقیق با زبان C# و با استفاده از نرم‌افزار Microsoft Visual studio 2019 پیاده‌سازی شده است. در این پیاده‌سازی روش پیشنهادی ۱۰۰ سند اصلی و ۲۰ سند مشکوک ذخیره شده است که در برخی از این اسناد ۱۰-۲۰ درصد، در برخی ۲۰-۴۰ درصد، در برخی ۴۰-۵۵ درصد و مانند آن کپی برداری شده است. سپس مطابق روش پیشنهادی ابتدا اسناد با فرمت PDF به XML تبدیل شده و در ادامه پس از پیش-پردازش و بلاک‌بندی مشابهت اسناد مشخص می‌شود. تمامی کدها و فایل‌های مورد استفاده در آدرس GitHub زیر قابل دسترسی می‌باشد.

https://github.com/negarmajma/Detection-of-plagiarism.git

برای مشخص شدن بهتر فایل‌های مورد استفاده در جدول (۲) نمونه‌ای از چند فایل اصلی که در آنها ۵۵ تا ۷۰ درصد تشابه ایجاد شده است، نشان داده می‌شود. نمونه‌های دیگر به همراه درصدهای متفاوت در آدرسی که در بخش کدها مقاله وجود دارد، قابل مشاهده است.

جدول (۲): نمونه‌ای از ایجاد اسناد کپی با درصد تشابه ۵۵ تا ۷۰

شماره فایل اصلی	تعداد کل کلمات بلاک	تعداد کارکتر	تعداد کلمات بلاک بعد از حذف Stop word	تعداد کارکتر بعد از حذف	تعداد کلمات کپی شده	تعداد کارکتر کپی شده	تعداد کلمات کپی شده بدون Stop word	تعداد کارکتر بدون Stop word	نسبت
۱	۱۵۴	۱۰۷۱	۹۵	۸۶۶	۱۰۷	۷۳۸	۶۵	۵۹۳	۰/۶۸
۶	۱۱۹	۷۷۰	۶۸	۵۸۴	۷۳	۴۸۳	۴۳	۳۷۰	۰/۶۳
۱۱	۱۹۸	۱۲۱۴	۱۰۸	۸۷۳	۱۲۷	۷۹۳	۷۱	۵۷۶	۰/۶۵
۱۸	۱۵۶	۱۰۰۷	۹۴	۷۸۴	۸۹	۵۶۶	۵۳	۴۴۳	۰/۵۶
۲۴	۲۱۷	۱۳۷۳	۱۲۴	۱۰۳۱	۱۳۸	۹۱۱	۸۰	۷۱۰	۰/۶۴
۲۵	۲۱۵	۱۲۰۴	۱۱۵	۸۳۸	۱۲۷	۷۱۹	۶۶	۵۰۹	۰/۵۷
۲۷	۴۳۸	۲۹۲۰	۲۷۶	۲۳۲۴	۳۰۸	۲۰۳۸	۱۹۱	۱۵۹۹	۰/۶۹

بازیابی با استفاده از یک فرم اجرایی انجام شده که فرم مذکور در شکل (۳) نمایش داده شده است. دقت داشته باشید که پس از اجرای مشابهت‌سنجی گزارش این مشابهت‌سنجی، که شامل قسمت‌های شناسایی شده به عنوان سرقت ادبی و بازدهی روش پیشنهادی (دقت، فرخوان، امتیاز اف) به صورت یک فایل گزارش ذخیره خواهد شد.

شکل (۳): فرم اجرایی روش پیشنهادی

۴. نتایج و بحث

در این بخش برای تایید دقت و بازدهی روش پیشنهادی، سه نوع آزمایش طراحی و اجرا شده است: در دسته اول آزمایشات معیار ارزیابی روش با تغییر دادن تعداد اسناد موجود در مخزن که سند مشکوک یا کپی شده از روی آنها به دست آمده، مشخص شده است؛ در دسته دوم آزمایشات تعداد جملاتی که از اسناد اصلی موجود در خزانه در سند مشکوک به کپی قرار گرفته‌اند به دست می‌آید و در نهایت در دسته سوم آزمایشات، با تغییر دادن تعداد اسناد موجود در پایگاه داده نتایج آزمایش بررسی می‌شود. این آزمایشات بر روی پایگاه داده‌ای ساختگی در حوزه مقالات علمی به زبان انگلیسی و بر روی سیستمی با مشخصات نشان داده شده در جدول (۳)، انجام شده است.

جدول (۳): سخت‌افزار استفاده شده

پردازنده	مقدار حافظه	سیستم عامل
اینتل i7	۱۶ گیگابایت	ویندوز ۱۰

۴،۱. معیارهای ارزیابی روش پیشنهادی

به منظور ارزیابی روش‌های تشخیص اصالت متون ادبی، از معیارهای Precision، Recall و Fmeasure استفاده می‌شود. اگر S مجموعه اسناد مشکوک به سرقت ادبی باشد و R مجموعه مواردی که پس از اجرای راه‌کار تشخیص اصالت متون ادبی بازیابی شده باشند، معیارهای Precision و Recall به صورت زیر تعریف می‌شوند:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \qquad (۸)$$

$$\text{Recall} = \frac{TP}{\text{TP} + \text{FN}} \qquad (۹)$$

که در آن TP تعداد اسنادی که همپوشانی دارند و به درستی همپوشانی آنها تشخیص داده شده است، FP اسنادی که همپوشانی ندارند ولی به اشتباه دارای همپوشانی تشخیص داده شده‌اند و FN اسنادی که همپوشانی ندارند و به درستی فاقد همپوشانی شناخته شده‌اند. همچنین Fmeasure که ترکیبی از دو معیار فوق می‌باشد و میانگین هارمونیک آنها است، به صورت رابطه‌ی تعریف می‌گردد:

$$Fmeasure = \frac{2 \times Recall \times Precision}{Recall + Precision} \qquad (۱۰)$$

۴،۲. مشخصات پایگاه داده

پایگاه داده‌ی ساختگی شامل ۱۲۰ مقاله با فرمت مقالات علمی است که از این بین ۲۰ سند به عنوان سند اصلی و ۲۰ سند به عنوان سند تقلب شده در نظر گرفته شده است. در این ۲۰ سند به مقادیر تصادفی تقلب صورت پذیرفته است.

۴،۳. ارزیابی روش پیشنهادی

آزمایش انجام شده در این پژوهش به سه دسته آزمایشات اصلی و در نهایت شناسایی بهترین پارامترهای پژوهش ختم خواهد شد. در بخش اول پارامترهای اصلی این روش شناسایی خواهند شد. مهمترین پارامتر روش پیشنهادی در بخش جمله‌بندی است که تعداد لغات تشکیل‌دهنده یک جمله چه عددی باشد. به طور طبیعی هرچه تعداد کلمات بیشتر شود، دقت سیستم در کپی محض افزایش می‌یابد. اما در مقابل بازنویسی یا خلاصه‌نویسی

متن به شدت آسیب‌پذیر می‌شود، یعنی در حالتی که پس از کپی کردن، نظم چینش لغات کپی شده تغییر یابد روش پیشنهادی با تعداد لغات زیاد در جمله آسیب‌پذیر خواهد بود. با این حال، روش پیشنهادی تنها کپی محض را مورد بررسی قرار داده است. جدول (۴)، به بررسی پارامتر تعداد لغات پرداخته است. در این آزمون ۱۰ سند حاوی ۱۰ پاراگراف دارای کپی در ۵ حالت مورد مقایسه قرار گرفته‌اند.

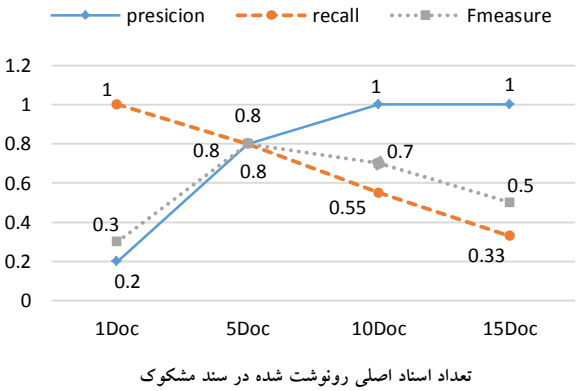
جدول (۴): آزمون تعداد لغات جمله

تعداد لغات	۱ لغت	۲ لغت	۳ لغت	۴ لغت	۵ لغت
	یک جمله				
دقت تشخیص	٪ ۸۰	٪ ۸۲	٪ ۸۸	٪ ۹۰	٪ ۹۴

همان‌گونه که از نتایج جدول (۴) قابل مشاهده است، با افزایش تعداد لغات در شناسایی کپی محض دقت نیز افزایش یافته است، اما در ابتدای این بخش نیز ذکر شد که در حالتی که متن کپی شده بازنویسی شود، افزایش تعداد لغات به شدت در افزایش دقت تاثیر خواهد داشت.

۴،۳،۱. دسته اول آزمایشات

در این آزمایش، تعداد سندهایی که سند مشکوک یا کپی شده از روی آنها کپی شده است، تغییر می‌یابد. این تغییرات از ۱ سند تا ۱۵ سند بوده و تاثیرات آن بر روی سه معیار ارزیابی پژوهش بررسی شده است. ارزیابی Precision، Recall و Fmeasure در تایید اصالت متون ارزیابی شده با استفاده از روش پیشنهادی در شکل (۴) نشان داده شده است.



شکل(۴): تاثیر تعداد اسناد اصلی در الگوریتم روش پیشنهادی

همان‌گونه که در شکل (۴) مشخص است، مقدار Fmeasure با ۵ سند اصلی در نقطه اوج نمودارها قرار دارد. تا پیش از ۵ سند و پس از آن روش پیشنهادی به کیفیت مورد نظر نرسیده است. بعد از این نقطه با توجه به افزایش تعداد اسناد اصلی و بالا رفتن احتمال خطا مقدار Fmeasure در هر دو حالت بعد کاهش یافته است. همچنین دقت با افزایش تعداد سندها به طور مستمر افزایش یافته و نرخ فرخوان نیز با افزایش تعداد سند کاهش داشته است. برای مشخص‌تر شدن شرایط این آزمایش یکی از حالاتی که برای ارزیابی اجرا شده است به تفصیل شرح داده می‌شود. در این نمونه، یک سند با ۱۰ سند دیگر مقایسه شده و میزان شباهت و شماره سندی که شباهت تشخیص داده شده در جدول (۵) نشان داده شده است.

جدول (۵): میزان شباهت یک سند نمونه با ده سند دیگر

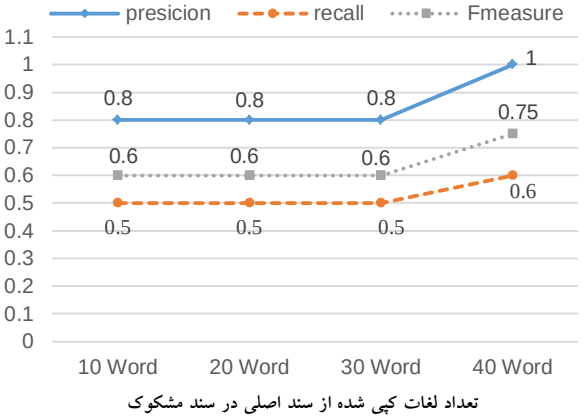
شماره سند که شباهت با آن تشخیص داده شده	درصد شباهت
۴۰	۰/۳۹
۷۰	۰/۳۳
۱۱۰	۰/۳۱
۸۷	۰/۳۰
۶۲	۰/۲۹
۴۳	۰/۲۵
۷۵	۰/۲۵
۳۰	۰/۲۴
۵۰	۰/۲۴
۲۰	۰/۲۳

اسنادی که در جدول (۵) با رنگ قرمز و توپر نشان داده شده‌اند، اسنادی هستند که با توجه به لغات مشابه، به اشتباه و نادرست، مشابه تشخیص داده شده‌اند. اکنون معیارهای دقت و نرخ فراخوانی و Fmeasure را در مورد این اسناد محاسبه می‌کنیم. برای مثال اگر $k = 1$ در نظر گرفته شود (به این معنی که تنها با یک سند مقایسه انجام می‌دهیم)، سند ردیف اول جدول (۵)، یک سند درست تشخیص داده شده است که میزان شباهت با سند اصلی مورد مقایسه را دارد؛ لذا دقت برابر ۱ و فراخوانی برابر ۰/۱ است ($TP = 1$، $FP = 0$ و $FN = 9$). در مثالی دیگر برای $k = 4$، مطابق جدول (۵)، سه سند به درستی و یک

سند نادرست، مشابه تشخیص داده شده است، لذا دقت برابر ۰/۷۵ و فراخوانی برابر ۰/۳ و Fmeasue برابر ۰/۴۲ است (وقتی که $TP = 3$، $FP = 1$ و $FN = 7$ است). سایر موارد به طور مشابه محاسبه می‌شوند.

۴،۳،۲. مجموعه آزمایش دوم

در این آزمایش تعداد لغت‌های کپی شده مد نظر است. بدین منظور باید تغییرات لغات در متن کپی شده بررسی و ارزیابی شود. در این بخش تعداد لغات مشخصی (از ۱۰ تا ۴۰ لغت) از ده سند اصلی کپی می‌شود تا عملکرد روش پیشنهادی در مواجه با این چالش مشخص شود. ارزیابی دقت، فرخوان و Fmeasure در کشف اصالت متون ادبی با استفاده از روش پیشنهادی در این حالت در شکل (۵) نشان داده شده است.

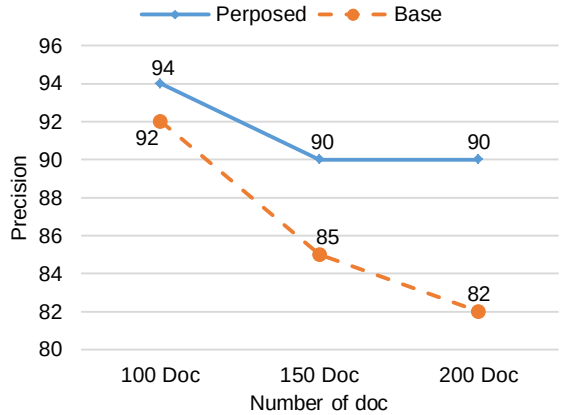


شکل (۵): تاثیر تعداد لغات بر روی نتایج روش پیشنهادی

ارزیابی دقت، فرخوان و Fmeasure در کشف سرقت ادبی نشان می‌دهد که عملکرد روش پیشنهادی برای حالتی که تعداد لغت‌ها کمتر از ۳۰ است، یکنواخت بوده و بازدهی ۸۰ درصدی در دقت کشف سرقت ادبی داشته است. اما در حالتی که تعداد لغات کپی شده به ۴۰ لغت افزایش یابد، دقت روش پیشنهادی افزایش یافته است.

۴،۳،۳. مجموعه آزمایش سوم

در این آزمایش با تغییر تعداد سندهای موجود در پایگاه داده معیارهای ارزیابی بررسی می‌شوند. به صورتی که تعداد اسناد موجود در خزانه هر بار با تعداد ۱۰۰، ۱۵۰ و ۲۰۰ سند مورد ارزیابی قرار می‌گیرد. نتایج این آزمون که به نوعی مقایسه روش پیشنهادی و روش مرجع [۱] است، در شکل (۶) نمایش داده شده است. روش پیشنهادی در حالتی که ۱۰۰ سند در خزانه جهت مقایسه باشد، دقت ۹۴ درصدی دارد، این در حالی است که روش مرجع [۱] در این حالت دارای دقت ۹۲ درصدی است. با افزایش تعداد سند در خزانه به میزان ۱۵۰ سند (یعنی افزایش ۵۰ درصدی تعداد اسناد)، دقت تشخیص رو به کاهش بوده است. روش پیشنهادی با ۱۵۰ سند در خزانه، ۹۰ درصد دقت تشخیص سرقت ادبی داشته، اما روش مرجع [۱] با کاهش ۷ درصدی، دقت ۸۵ درصدی در تشخیص سرقت ادبی را نشان می‌دهد که این امر نشان‌دهنده کمتر بودن افت دقت تشخیص در روش پیشنهادی است. در ادامه، تعداد اسناد خزانه به ۲۰۰ سند افزایش یافته است که در این حالت روش پیشنهادی هیچ کاهش دقتی نداشته است، اما روش مرجع [۱] باز هم با افت دقت تشخیص سرقت ادبی همراه بوده و دقت آن در این حالت ۸۲ درصد بوده است. روش پیشنهادی به غیر از این که توانسته دقت بالاتری در تشخیص سرقت ادبی داشته باشد، پایداری قابل قبول‌تری به نسبت روش مرجع [۱] در مقابل چالش افزایش تعداد اسناد موجود در خزانه دارد.



شکل (۶) تاثیر تعداد اسناد موجود در پایگاه داده

۵. ارزیابی

در ادامه، در سه بخش به صورت تفصیلی به ارزیابی نتایج روش پیشنهادی پرداخته خواهد شد.

ارزیابی تاثیر تعداد اسناد اصلی: با افزایش تعداد اسناد اصلی دقت روش پیشنهادی افزایش داشته اما فراخوان و امتیاز اف روش پیشنهادی با کاهش رو به رو بوده است. این موضوع نشان می‌دهد که روش پیشنهادی با افزایش تعداد اسناد سرقت ادبی‌های انجام شده را با دقت بیشتری شناسایی می‌کند، اما دقت در شناسایی مواردی که سرقت ادبی ندارند، کاسته می‌شود. به عبارت دیگر، مواردی که سرقت ادبی نیستند، اما مدل سرقت تشخیص می‌دهد رو به افزایش خواهد بود.

ارزیابی تاثیر تعداد لغات بر روی نتایج: در نتایج مشاهده شد که با افزایش تعداد لغات در جمله‌بندی، دقت روش پیشنهادی افزایش داشته که دلیل این موضوع کاملا روش است. با افزایش تعداد لغات، جملات به دست آمده موضوعیت مشخص‌تری را نمایش می‌دهند و این موضوع باعث می‌شود که اسناد بازیابی شده برای مرحله دوم شناسایی، دقیق‌تر انتخاب شوند.

بررسی تاثیر تعداد اسناد موجود در پایگاه داده: با افزایش تعداد اسناد تعداد خطاهای مدل پیشنهادی افزایش یافته است. این موضوع دور از ذهن نبوده و به دلیل افزایش تعداد مقایسات دقت نیز کاسته شده است، اما نکته مثبت در این موضوع کاهش ناچیز دقت به نسبت افزایش تعداد اسناد بوده است که به نوعی نقطه قوت برای مدل پیشنهادی محسوب می‌شود.

۶. نتیجه‌گیری و کارهای آتی

مقاله جاری، روشی برای تشخیص سرقت ادبی در متون علمی پیشنهاد داده است که با انجام سه سری آزمایشات مختلف بر روی اسناد ساختگی مورد ارزیابی قرار گرفته و در نهایت نتایج حاصل از این آزمایشات با روش مرجع [۱] مقایسه شد. در سری آزمایشات اول، تاثیر تعداد اسناد موجود در مخزن که رونوشت غیرمجاز از روی آنها انجام گرفته بر روی تشخیص شباهت اسناد بررسی گردید که در این آزمایشات اغلب نتایج Fmeasure در روش پیشنهادی بهبود داشت به صورتی که دقت بعد از ۱۰ سند به ۱۰۰ درصد رسیده است. در سری آزمایش دوم، تاثیر تعداد لغات رونوشت شده از روی اسناد موجود در سند مشکوک بررسی گردید. نتایج نشان داد که عملکرد روش پیشنهادی تا پیش از ۳۰ لغت در حد متوسط بوده ولی با ۴۰ لغت به دقت ۱۰۰ درصدی رسیده است. در آخر یک آزمایش مقایسه‌ای با روش مرجع مبتنی بر تعداد اسناد موجود در مخزن انجام شد که روش پیشنهادی دقت بهتری به نسبت روش مرجع داشته و در بهترین حالت به نسبت روش مرجع [۱] بهبود ۵ درصدی را نشان می‌دهد.

برای ادامه کار این مقاله، بررسی انواع معیارهای شباهت مانند فاصله اقلیدوسی و فاصله همیلتن و مقایسه آنها با معیار شباهت کسینوسی مورد استفاده در این مقاله و همچنین بررسی اسناد با لغات رونوشت بیشتر، پیشنهاد می‌شود.

تعارض منافع: نویسندگان اعلام می‌کنند که هیچ تعارض منافعی ندارند.

مراجع

[1] G. Sarwar, C. O'Riordan, and J. Newell, "Passage Level Evidence for Effective Document Level Retrieval," Proc. 9th Int. Joint Conf.Knowl. Discov. Knowl. Eng. Knowl. Manag. (KDIR), pp. 83-90, 2017, doi: 10.5220/0006502800830090.

[2] I. Jaric, "High Time for a common Plagiarism Detection System," Scientometrics, vol. 106, no. 1, pp. 457-459, 2016, doi: 10.1007/s11192-015-1756-6.

[3] I. Masic, E. Begic, and A. Dobraca, "Plagiarism Detection by Online Solutions," Informatics Empowers Healthcare Transformation, ICIMTH 2017, Athens, Greece, pp. 227-230, 2017, doi: 10.3233/978-1-61499-781-8-227.

[4] V. Kanjirangat and D. Gupta, "Study on Extrinsic Text Plagiarism Detection Techniques and Tools," J. Eng. Sci. Technol. Rev., vol. 9, no. 5, pp. 8-22, 2016, doi: 10.25103/jestr.095.02.

[5] S.M. Alzahrani, N. Salim, and A. Abraham, "Understanding Plagiarism Linguistic Patterns, Textual Features, and Detection Methods," IEEE

Trans. Sys. Man Cybern., Part C, vol. 42, no. 2, pp. 133-14, 2012, doi: 10.1109/TSMCC.2011.2134847.

[6] M. Potthast, A. Barron-Cedeno, B. Stein, and P. Rosso, "Cross-language Plagiarism Detection," Lang. Resour. Evaluation, vol. 45, no. 1, pp. 45-62, 2011, doi: 10.1007/s10579-009-9114-z.

[7] C. Grozea, C. Gehl, and M. Popescu, "ENCOPLOT: Pairwise Sequence Matching in Linear Time Applied to Plagiarism Detection," 3rd PAN Workshop Uncovering Plagiarism, Authorship and Social Software Misuse, WUPAS, pp. 10-18, 2009.

[8] C. Basile, D. Benedetto, E. Caglioti, G. Cristadoro, and M. Esposti, "A Plagiarism Detection Procedure in Three Steps: Selection, Matches and Squares," Proc. SEPLN, pp. 19-23, 2009.

[9] M. Elhadi and A. Al-Tobi, "Duplicate Detection in Documents and Webpages using Improved Longest Common Subsequence and Documents Syntactical Structures," 4th Int. Conf. Comput. Sci. Convergence Inf. Technol., pp. 679-684, 2009, doi: 10.1109/ICCIT.2009.235.

[10] K. Koroutchev and M. Cebrian, "Detecting Translations of the Same Text and Data with Common Source," J. Stat. Mech. Theory Exp., vol. 10, 2006, doi: 10.1088/1742-5468/2006/10/P10009.

[11] Y. Li, D. McLean, Z.A. Bandar, J.D. O'shea, and K. Crockett, "Sentence Similarity Based on Semantic Nets and Corpus Statistics," IEEE trans. Knowl. Data Eng., vol. 18, no.8, pp.1138-1150, 2006, doi: 10.1109/TKDE.2006.130.

[12] S. Gruner and S. Naven, "Tool Support for Plagiarism Detection in Text Documents," Proc. ACM SAC, Santa Fe, New Mexico, USA, pp. 776-781, 2005, doi: 10.1145/1066677.1066854.

[13] S. Alzahrani and N. Salim, "Statement-based Fuzzy-set IR Versus Fingerprints Matching for Plagiarism Detection in Arabic Documents," Proceedings 5th Postgraduate Annual Research Seminar (RARS 09), pp. 267-268, 2009.

[14] A.H. Osman, N. Salim, and A. Abuobieda, "Survey of Text Plagiarism Detection," Comput. Eng. Appl. J., vol. 1, no. 1, pp. 37-45, 2012, doi: 10.18495/comengapp.v1i1.5.

[15] T.W. Chow and M. Rahman, "Multilayer SOM with Tree-Structured Data for Efficient Document Retrieval and Plagiarism Detection," IEEE Trans. Neural Networks, vol. 20, no. 9, pp. 1385-1402, 2009, doi: 10.1109/TNN.2009.2023394.

[16] S. Rakian, E.F. Safi, and H. Rastegari, "A Persian Fuzzy Plagiarism detection Approach," J. Inf. Sys. Telecommun., vol. 3, no. 11 pp. 182-190, 2015, doi: 10.7508/jist.2015.03.007.

[17] D. Gupta, K. Vani, and L. Leema, "Plagiarism Detection in Text Documents using Sentence Bounded Stop word N-Grams," J. Eng. Sci. Technol., vol. 11, no. 10, pp. 1403-1420, 2016.

[18] M.A. Sanchez-Perez, A. Gelbukh, G. Sidorov, and H. Gomez-Adorno, "Plagiarism Detection with Genetic-based Parameter Tuning," Int. J. Pattern Recognit. Artif. Intell., vol. 32, no, 1, 2018, doi: 10.1142/S0218001418600066.

[19] M. Schubotz, O. Teschke, V. Stange, N. Meuschke, and B. Gipp, "Forms of Plagiarism in Digital Mathematical Libraries," Int. Conf. Intell. Comput. Math., pp. 258-274, 2019, doi: 10.1007/978-3-030-23250-4\18.

[20] C.-Y. Chang, S.-J. Lee, C.-H. Wu, C.-F. Liu, and C.-K. Liu, "Using Word Semantic Concepts for Plagiarism Detection in Text Documents," Inf. Retr. J., vol. 24, no. 4, pp. 298-321, 2021, doi: 10.1007/s10791-021-09394-4.

[21] S.A. Asgari, M. Enayati, G. Abaei, and M Binesh-Mavesti, "Providing an Improved Webmining Algorithm for Semantic Web," Soft Comput. J., vol. 5, no. 1, pp. 2-13, 2016 [In Persian].

[22] Z. Farahmandpoor, H. Nikmehr, M. Mansoorizade, and O. Tabibzadeh-Ghamsary, "A Novel Intelligent Persian Authorship System based on Writing Style," Soft Comput. J., vol. 1, no. 2, pp. 26-35, 2012, dor: 20.1001.1.23223707.1391.1.2.60.9 [In Persian].

[23] E.M.B. Nagoudi, H. Cherroun, and A. Alshehri, "Disguised Plagiarism Detection in Arabic Text documents," 2nd Int. Conf. Natural Lang. Speech Process., pp. 1-6, 2018, doi: 10.1109/ICNLSP.2018.8374395.

[24] M. Agrawal and D.K. Sharma, "A State of Art on Source Code Plagiarism Detection," 2nd Int. Conf. Next Gener. Comput. Technol., pp. 236-241, 2020, doi: 10.1109/NGCT.2016.7877421.

[25] H. Arabi and M. Akbari, "Improving plagiarism detection in text document using hybrid weighted similarity," Expert Sys. Appl., vol. 207, 2022, doi: 10.1016/j.eswa.2022.118034.

[26] A. Ali and A.Y. Taqa, "Analytical Study of Traditional and Intelligent Textual Plagiarism Detection Approaches," J. Educ. Sci., vol. 31, no. 1, pp. 8-25, 2022, doi: 10.33899/edusj.2021.131895.1192.

[27] S. Zangoei, "Examination of Authors' Stylistic Elements of Electronic Messages based on Researched Studies," Soft Comput. J., vol. 6, no. 2, pp. 60-71, 2017, dor: 20.1001.1.23223707.1396.6.2.5.9.